HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT

BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 372

FESTUNG METS –
WIE STAND-OFF-ANNOTATIONEN IN RDF DYNAMISCHE DOKUMENTSTRUKTUREN AUS XML-HIERARCHIEN BEFREIEN KÖNNTEN

EINE UNTERSUCHUNG AM DATENMANAGEMENT VON „SCRIPTA PAEDAGOGICA ONLINE“

VON
MARTIN WÜNSCH

FESTUNG METS –
WIE STAND-OFF-ANNOTATIONEN IN RDF DYNAMISCHE DOKUMENTSTRUKTUREN AUS XML-HIERARCHIEN BEFREIEN KÖNNTEN

EINE UNTERSUCHUNG AM DATENMANAGEMENT VON „SCRIPTA PAEDAGOGICA ONLINE“

VON
MARTIN WÜNSCH

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 372

Wünsch, Martin

Festung METS – Wie Stand-off-Annotationen in RDF dynamische Dokumentstrukturen aus XML-Hierarchien befreien könnten : Eine Untersuchung am Datenmanagement von „Scripta Paedagogica Online“ / von Martin Wünsch. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2014. – 86 S. : graph. Darst. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 372)

ISSN 14 38-76 62

Abstract:

Die technischen Lösungen zu einer beschreibungssprachlichen Uniformlösung für die Dissemination von Ergebnissen der Massendigitalisierung im nationalen Maßstab beinhalten Schwachstellen. In dieser Arbeit soll der Schwachpunkt bei der Seiten-URN im Open Source Projekt „Goobi.production“ dargestellt werden. Die sprachlichen Restriktionen von XML stellen die METS-Dokumente gegenüber neueren Beschreibungssprachen des Semantic Web als Hemmnis dar, wenn diese Dokumente als dynamische Objekte genutzt werden müssen. Die Arbeit behandelt die Erweiterung der Algorithmen mit Techniken des Semantic Web vor einem aufgespannten Problemhorizont.

Diese Veröffentlichung geht zurück auf eine Masterarbeit im postgradualen Fernstudiengang M. A. Bibliotheks- und Informationswissenschaft (Library and Information Science) an der Humboldt-Universität zu Berlin.

Online-Version: http://edoc.hu-berlin.de/series/berliner-handreichungen/2014-372

Inhaltsverzeichnis

Technische Anmerkung:

Im PDF-Dokument sind alle Querverweise und Quellennachweise mit Hyperlinks zum jeweiligen Pendant an anderen Dokumentstellen ausgestattet. Die Hyperlinks besitzen aus ästhetischen Gründen einen wie üblich unsichtbaren Rahmen. Die PDF-Lesezeichen geben das Inhaltsverzeichnis wieder zuzüglich weiterer Einträge zur Verbesserung der Navigation im Dokument.

Danksagung

Ich danke dem Deutschen Institut für Internationale Pädagogische Forschung für die großzügige Unterstützung und Ermöglichung bester Gelingensbedingungen für das Studium und für diese Handreichung.

1. Einleitung

Diese Arbeit stellt einen Versuch dar, die Techniken und Modelle von Wissensrepräsentation eines digitalbibliothekarischen Produktions- und Präsentationssystems kritisch zu betrachten. Die Arbeit wird hierzu von Erscheinungen am Untersuchungsobjekt berichten. Mit der Analyse der Phänomene wird versucht, das Wesenhafte, also die invarianten Bestimmungen der Sachverhalte, zu erfassen. Daraufhin können (bezogen auf den gefilterten Phänomenbereich) metasprachliche Fakten formuliert werden, die für die Untersuchung dienlich und plausibel sind. Schließlich können Hypothesen diese Fakten in einen Modus bringen, der zur Diskussion einlädt.

Die Arbeit liefert in **Kapitel 2** eine bisher noch nicht verfügbare Zusammenstellung von den wissenschaftsregulierenden Bedingungen und informationstechnischen Normierungen die konkret auf das Untersuchungsobjekt Einfluß nehmen, aber auch dessen Entstehung geprägt haben. Bei dieser Reflexion ist vor allen die Überlegung motivierend, daß den Anwendern und Interessenten des zu untersuchenden Produktions- und Präsentationswerkzeugs einer digitalen Bibliothek[1] dieser Überblick so noch nicht angeboten wird. In **Kapitel 3** wird eine ebenso neuartige Zusammenstellung von anwenderbezogenen Informationen aus dem Material der „Bauanleitungen“ bzw. „Handbüchern“[2] des Systems vorgestellt. Nur in diesem beiden Kapiteln 2 und 3 befinden sich auch bewertende Fazits und Kritiken in hervorgehobenen Textblöcken. Im **Kapitel 4** werden anhand eines Fallbeispiels die Probleme der Systemanwendung wegen einer funktionalen Haltelinie im System dargestellt. Der Praxisbericht behandelt eine typische, nicht zu seltene Fehlersituation im Workflow einer digitalen Bibliothek, die einen Anlaß zur Systemkorrektur darstellt. Aus dem aufzuzeigenden „toten Winkel“ des behandelten

[1] Das Adjektiv „digitale“ wird im Text nicht wie eine Wortmarke groß geschrieben, sondern als normales Wort verwendet. Die rhetorische Hervorhebung „Digitale Bibliothek“ ist dennoch frei nach Belieben assoziierbar.

[2] Ein umfassende oder zumindest zentrale und didaktisch aufbereitete Dokumentation gibt es nicht. Ein Überblick über die vertreuten Informationsquellen und deren Einzelbewertungen müssen aufgrund des thematischen Schwerpunkts dieser Arbeit zurückgestellt werden.

Dokumentenmodells soll ein neu eingeführtes Diskursobjekt im Kontext einer Kontrolle von Quellenauthentizität herausführen. In **Kapitel 5** werden aus der Perspektive einer modelltheoretischen Semantik die Grenzen des XML-Dokumentenmodells dargestellt. Mit dem in Kapitel 4 aussagenlogisch formulierten Diskursobjekts wird beispielhaft eine prädikatenlogische Schlußfolgerung dargestellt, um damit die Unterscheidung zwischen XML-Parser und Reasoner [00a] verständlich zu machen. In **Kapitel 6** werden mit dem in Kapitel 5.2.1 eingeführten "Digiproblem"-Konzept die Anforderungen an ein erweitertes System bzw. an eine erweiterte Metasprache diskutiert und ein Werkzeug genannt, das für die Einführung von Semantic Web Technologien motivierend ist. Schließlich sollen in **Kapitel 7** einige über diesen Arbeitsrahmen hinausgehenden Wünsche und Lasten angesprochen werden.

Der Verfasser hat keine programmiertechnischen Kenntnisse, noch ist er examinierter Informatiker oder Mathematiker. Ein detailliert informatisches Analysieren des Untersuchungsobjekts, wofür Java-programmierte Quellcodes [00b] gelesen werden müßten, ist in dieser Arbeit entsprechend ausgeschlossen. Ebenso kann sich der Verfasser (noch) nicht als aktiver Mitgestalter konkreter Softwareprojekte mit dem Einsatz von Semantic Web Technologien und deren Werkzeugen bezeichnen. Das postgraduale Fernstudium der Bibliotheks- und Informationswissenschaft und diese Abschlußarbeit sind bewußt als Mittel zur Initialzündung für solche Aktivitäten ausgewählt worden. Als Anwender der zu untersuchenden Programme für den Aufbau und Betrieb einer digitalen Bibliothek sollen hier die Erfahrungen genutzt sein, um mit und nach einem strukturierten Analyseprozeß Erweiterungsvorschläge abgegeben zu können.

2. Zum Kontext des Untersuchungsobjekts

Ungefähr 40 Bibliotheken in Europa steuern zu Beginn des Jahres 2013 über ein Workflowmanagement für die Digitalisierung von gedruckten Beständen ihre Produktionen in digitale Repositorien über die Applikation „Goobi.Production“ [00c] und publizieren die Konstrukte über „Goobi.Presentation“ [00d] (open source, Typo3) oder den „intranda viewer“ [00e] (proprietär, Java). Auf den Informationsseiten dieses open source Applikationsprojekts „Goobi“ werden 26 dieser Bibliotheken namentlich aufgelistet [00f]. Das digitale Angebot verdankt in erster Linie seine Erstellung aus DFG-geförderten retrospektiven Digitalisierungsprojekten, die nach entsprechenden Praxisregeln [00g] ausgerichtet sein müssen. Diese empfohlenen Praxisregeln werden de facto als Richtlinien aufgefaßt. Die Verfasser des Papiers betonen zwar, daß hier lediglich eine Absicht besteht, den „Antragstellern die Planung von Digitalisierungsprojekten zu erleichtern und die Begutachtung von Anträgen vergleichbar zu gestalten“ und daß sie nicht das Ziel verfolgen, „Hürden aufbauen“. Es bleibt jedoch zu vermerken, daß „durch die Formulierung von Standards“ [00g , S. 4] in der Konsequenz „(j)ede Abweichung von diesen Standards [...] ausführlich begründet werden (muß).“ [00g , S. 45] Das Papier ist dem normativen Charakter entsprechend auf der DFG-Website unter der Veröffentlichungsrubrik „Richtlinien“ als verfügbare elektronische Ressource plaziert. [00h]

2.1 DFG-Praxisregeln und Technische Parameter

In dem Kapitel „Technische Parameter der digitalen Reproduktion“ erläutern die Praxisregeln[3] mit 6.640 Wörtern[4] detailliert die signaltechnischen und physikalischen Berücksichtigungen für ein optimales Digitalisierungsergebnis mit Hilfe bildgebender Geräte und technischer Sorgfalt. Im Text wird zusätzlich auf das deutsche Portal der

[3] Es wird stets auf die DFG-Praxisregeln „Digitalisierung“ (Stand: 02/2013) = [00g] Bezug genommen.

[4] Es erfolgte eine einfache Wortzählung: (a) PDF-Text markiert und (b) Kopieren in einen Texteditor, (c) Kopfzeilentexte entfernt, (d) Zählung des Textes mit Hilfe von Microsoft Word.

Langzeitarchivierung verwiesen, über das potentiell das 634-seitige nestor-Handbuch [NOSS00, S.3] erreichbar ist. Ebenfalls verweist eine weitere Fußnote im Kapitel auf die Schrift „Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files, August 2010.“ der Federal Agencies Digitization Guidelines Initiative (FADGI). [00i] Es soll für den nachfolgenden Vergleich lediglich der Grad der Ausführlichkeit von technischen Erklärungen zur Kenntnis genommen sein.

2.2 DFG-Praxisregeln und Metadaten METS / MODS

Die in den Praxisregeln enthaltenen Empfehlungen zu den Metadaten umfassen im Vergleich zum technischen Abschnitt im Hauptteil 1.191 Wörter. Im zugehörigen Dokument „Anhang A: METS/MODS-Profil für die Darstellung im DFG-Viewer und Übermittlung per OAI“ [Deut13] umfassen 419 Wörter den Fließtext. Es wird zusätzlich das 28-teilige MODS-DFG-Standard-Set für Printbestände tabellarisch und kommentiert aufgeführt. Dieses Set wird mit folgendem Text eingeführt: „Der MODS-Standard bietet ein vereinfachtes subset zu MARC 21, daher sollten automatische Konversionen aus gängigen Katalogen leicht möglich sein. Für die Anzeige im Viewer sind nur wenige Pflichtfelder erforderlich [...].“

Für diese standardisierte Metadatenreduktion gibt [Chen12] ein Beispiel für bibliographischen Metadaten einer Leichenpredigt von 1656. [00j] Die Metadatenreduktion, die von Chen vorgeführt wird, stellt das PICA+-Set dem MODS-Set als Migrationsquelle gegenüber. Es ist jedoch offensichtlich der Fall, daß die Formate PICA+ und MARC 21 in Bezug auf ihre Elemente jeweils die größere Menge als der MODS-Standard abbilden und somit das gleiche Szenario auf das DFG-Dokument übertragen werden kann. Der MODS-Standard transportiert jedoch ein ausgesprochen gewollt reduziertes bibliographisches Elementeset, das statt den numerisch kodierten Feldbezeichnern von MARC 21 natürlichsprachlich angelehnte Kunstwörter als kontrolliertes Vokabular benutzt. Jedoch ist MODS reicher bestückt als Dublin Core und dient einer „variety of purposes, and particularly for library applications.“ [00k] [5]

[5] Eine sehr hilfreiche Konkordanz zwischen den granularen Formaten USMARC, UKMARC, UNIMARC, Pica 3, Pica+, MAB1, MAB2, BIS-LOK, ZDB-Zeta und „allegro“ steht über die „allegro-Formatedatenbank“ frei - online und offline - zur Verfügung. [00l]

Kritik 1

Aus dieser vorerst oberflächlichen Beobachtung können dennoch folgende vorsichtigen Schlußfolgerungen aus den DFG-Praxisregeln gezogen werden, die die Kanalisierung von Algorithmen verschiedener Projektebenen im Förderbereich der Wissenschaftlichen Literaturversorgungs- und Informationssysteme (LIS) durch relativ schmale Standard-metadatenkorridore verdeutlichen:

a) Technische Aspekte werden aufwendiger dargestellt als Metadatensachverhalte. Die detaillierten Erklärungen tragen dem möglicherweise verbreiteten Umstand Rechnung, daß fachberufliche Arbeitskräfte für bildgebende Geräte seltener in Bibliotheken anzutreffen sind. Die Erklärungen im Text sollen auch eine Kritikfähigkeit vermitteln, um den Angeboten der externen kommerziellen Dienstleister als sachkompetenter Kunde begegnen zu können. Die Digitalisierungsprojekte sind im Falle von Erstanschaffungen von oder Modernisierungen der Infrastruktur mit relativ kurzen Produktzyklen und großer Produktvielfalt im Marktgeschehen konfrontiert, wobei eine rekursive Marktbeobachtung vermutlich selten „zwischen den Projekten“ vorgenommen wird[6]. Für den technischen Textmigrationspfad über OCR-Softwareprodukte (optisch-elektronische Schrifterkennung) wird als Herausforderung „der Markt (der sich) anbieterseitig dynamisch weiterentwickelt“ sogar wörtlich genannt. [00g , S. 32]

b) Bei den textlich kürzer behandelten Metadaten verweisen die Praxisregeln auf „einschlägige Spartenstandards“ bzw. „materialspezifische Standards“. [00g , S. 27] Im Fall von „gedruckten Textwerken“ ist dies „METS/MODS“ (ebd. S. 27). Es wird davon ausgegangen, „daß die der Digitalisierung zu Grunde liegenden analogen Objekte bereits primär in anerkannten digitalen Nachweissystemen erschlossen sind bzw. mit der Digitalisierung einhergehend erschlossen werden.“ Mit der Wahl des XML-basierten Standard METS/MODS versichert sich die Leitlinie, daß die „Daten in einem herstellerunabhängigen, sowohl semantischen als auch technischen Standardformat“ vorhanden sind. [00g , S. 25]

Die „Überlegungen [der DFG] beziehen sich ausschließlich auf deskriptive und struktu-

[6] Es gibt im Bereich bildgebender Techniken für das Digitalisieren von Bibliotheksgut kein nationales Kompetenzzentrum. Als ein solches Beispiel sei das Kompetenzzentrum für Videokonferenzdienste (VCC) an der TU Dresden zu nennen, das der Deutsche Forschungsverein e.V. schon seit 2002 betreibt. Es werden dort fortlaufend die Erfahrungen von Produkttests weitergegeben und Praxisworkshops angeboten. [09]

relle Metadaten.“ So sind Empfehlungen für „administrative (z.B. Rechteverwaltung) und technische (z.B. Dateitypen) Metadaten“ anderweitig einzuholen. [00g , S. 25]

Im Gegensatz zu der eher beschleunigten Entwicklung im Bereich der proprietären industriellen Infrastruktur (wie z.B. Kameras, Scanner, Buchscanner, chipbasierte Verarbeitungs- und Bilderstellungsalgorithmen) kann der gemeinfreie logische Systemunterbau in eine weniger beschleunigte nachhaltigere Verfahrensbeständigkeit geführt werden. Diese Entschleunigung zeigt sich jedoch auch bei einer ins Leere laufenden Vorgabe „METS/TEI für Handschriften (s. Anhang B)“. Diesen Anhang B gibt es zum Zeitpunkt der Recherche Anfang April 2013 (immer noch) nicht. Die METS/TEI-Kombination scheint aber verplant zu sein, da dieser Standard ebenfalls beim Kompetenzzentrum für Interoperable Metadaten (KIM) in der Gruppe „Digitalisierungsmetadaten“ als nichtverlinkte bzw. nichtreferenzierte Nennung in „Aktuelle Themen“ aufgelistet ist: „Mittelalterliche und frühneuzeitliche Handschriften (METS/TEI)“ Der Eintrag befindet sich in direkter Nachbarschaft zu „Drucke, Zeitschriften und Zeitungen (METS/MODS)“ und „Nachlässe und Autographen (METS/EAD)“. [00m] Der METS/MODS-Hyperlink zeigt direkt und kommentarlos auf die Startseite des DFG-Viewers.

Eine zentrale algorithmische Prüfinstanz zur verpflichteten Validierung der Projektergebnisse von digitalisierten gedruckten Textwerken ist der soeben genannte „DFG-Viewer“. [00n] Die Richtlinien „(empfehlen) nach dem derzeitigen Stand bei alten Drucken eine Orientierung an **METS** oder **TEI**. Dessen ungeachtet soll in jedem Fall der **DFG-Viewer** unterstützt werden, der auf METS beruht.“ [00g , S. 25] Diese - metaphorisch gesprochen - monophage Organisation des Viewers über eine METS/MODS-Kombination wird auch im PDF-Dokument [Funk09] und in der XML-Version zum „zvdd/DFG-Viewer METS-Profil – Version 2.0" [EnFu00a] genannt: „Sowohl der DFG-Viewer als auch das zvdd-Portal unterstützten lediglich deskriptive Metadatensektionen vom Typ MODS.“

Fazit 1 • Man kann aus dieser Gegenüberstellung technischer und sprachlicher Werkzeuge und Methoden folgende Arbeitsteilung, was die Erzielung einer Vollständigkeit der Online-Präsentation betrifft, ableiten:

- Die bestmögliche bildgebende Technik für gedruckte Bestände soll den bisweilen nicht wiederholbaren Digitalisierungsprozeß in einer weitgehend vollständige Abbildung der Überlieferungsform zum Ergebnis kommen lassen.
- Eine deutlich kleine Teilmenge der deskriptiven Metadaten dienen in den Containerobjekten hinreichend zur Identifizierung der Werke und unterstützen bei der elektronischen Präsentation einen simulierten bzw. virtuellen Zugang zur Überlieferungsform.
- Die originaltextlichen Inhalte der Überlieferungsformen sollen ebenfalls mit bestmöglichen Verfahren bei allenfalls entsprechender Quellenlage in elektronischen Volltext überführt werden.

Sodann man das Vollständigkeitspostulat auf den Überbau „DFG-Förderbereich der Wissenschaftlichen Literaturversorgungs- und Informationssysteme (LIS)“ überträgt und die dazugehörigen Aufwendungen in ihrer publizierten Größenordnung bemißt, welche da sind, „Erschließung und Digitalisierung handschriftlicher und gedruckter Überlieferung“: 14,8 Mio. EUR (2011), 13,8 Mio. EUR (2010), 21,4 Mio. EUR (2009)[7], sind die offenen und versteckten Haltelinien im Gesamtsystem immerhin von einer nicht unerheblichen Anzahl von Beteiligten schon länger erfahrbar. Es liegt somit der Gedanke nicht fern, daß die im Folgenden aufzuzeigenden problematischen Modellkompromisse ähnliche konservative Kraft besitzen, wie ein beispielhaftes Gegenteil, das John Unsworth [00o] aus den USA vorstellt. [Unsw10]

Das EVIA Digital Archive Project (2001-2009) - mit 4 Mio. US-Dollar Finanzierung ausgestattet - war „unique in its combination of preservation, annotation, and scholarly

[7] http://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/dfg_jb2011.pdf - abgerufen am 10.05.2013

http://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/dfg_jb2010.pdf - abgerufen am 10.05.2013

http://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/dfg_jb2009.pdf - abgerufen am 10.05.2013

publishing.“ [Burd10] Es wurden Videomagnetbandunikate mit ethnographischen Filmmaterial digitalisiert und annotiert, die akut durch Entmagnetisierung bedroht waren. Die drei zu unterscheidenden Projektbereiche Applikationsentwicklung, Annotation mit peer-review-Verfahren und Digitalisierung bildeten drei Interessenlager. Unsworth legt dar, daß das Material bei bestehender Durchsatzgeschwindigkeit noch 118 Jahre bis zum Finalisieren benötigen würde und erläutert, daß die Gründe für die dominante Applikationsentwicklung (was den Finanzbedarf betrifft) und Annotationsorganisation (was den Zeitaufwand betrifft) gegenüber der Digitalisierung in der Liebe zum Detail und zur Perfektion bei den Applikationen und in dem aufwendigen und verzögernden peer-review-Verfahren liegen und zudem noch der Umstand hinzukommt: „(O)nce you put that tool [die Annotationsumgebung, der Verf.] in the hands of the ethnographers, it turns out they have lots and lots to say, so much, in fact, that they can't justify their level of effort unless the results count for tenure and promotion, so now you need to implement a peer-review and publishing process in order to provide them with professional credit.“ [Unsw10]

Mit einem Blick zurück auf die deutsche Situation ist nach diesem Vergleich festzustellen, daß statistische Daten über das quantitative Ergebnis aller finanzierten Digitalisierungsprojekte nicht verfügbar sind. Eine Evaluation dieses DFG-Förderschwerpunkts fand zuletzt 2005 statt. [AlMa05] Das Dokument wird im April 2014 nach wie vor als aktuellste Wahl bei den Evalutionsstudien der DFG [00p] aufgeführt. Es kann nun aus den DFG-Praxisregeln und aus dieser Studie der Universität zu Köln eine vermutliche Paarung von Empfehlungen behauptet werden. Die Paarung bereitet der Studientext von 2005 über eine der „vier Grundhaltungen“ zur Erschließungstiefe der Quellen vor. Folgendes wurde aus den Aussagen der befragten Projekte klassifiziert: „In einer weiteren Gruppe von Projekten wird davon ausgegangen, daß die digitalisierte Information sozusagen eine Erweiterung des klassischen Katalogisats (in der vollen Bandbreite des Begriffes) ist. In solchen Fällen steht die Umsetzung der bisher bekannten Katalogstrukturen im Vordergrund, die durch den sekundären Zugriff auf das erschlossene Material in digitaler Form ergänzt werden.“ [AlMa05, S.28]

Die paarige Verwandtschaft zwischen der aktuellen DFG-Richtlinie und den pragmatischen Analysen der Evaluation zeigt sich in der „vorläufigen Empfehlung“ aus dem Gutachten:

„(a) Es sollte ein Minimalumfang an Metainformation festgeschrieben werden, der durch Retrodigitalisierungsprojekte erbracht werden muß. Dieser Minimalstandard sollte sich allerdings nicht an abstrakten Standardisierungsdiskussionen mit dem Ziel der Entwicklung möglichst umfassender Standards orientieren, sondern an dem pragmatischen Bedürfnis des Datentausches zwischen einzelnen Ressourcen. Hier wäre etwa die Forderung vorstellbar, daß das OAI-Protokoll auf der Basis des 'unqualified DC' zu unterstützen wäre.“ [AlMa05, S.29]

Im zweiten Teil der Empfehlung zu den Metadaten wird diese kleinste gemeinsame Metadatenkonvention als explizite Haltelinie für die Erschließungs- und Annotationstiefe in der finanzierten Projektphase genannt und als Argument für eine einfachere und gerechtere Antragsbegutachtung verwendet.

„(b) Solange die volle Bandbreite der oben beschriebenen Paradigmen für die Bereitstellung digitaler Ressourcen unterschiedslos akzeptiert wird, ist es praktisch nicht möglich, die Angemessenheit des Aufwands für und die Effektivität bei der Erhebung von Metadaten zu beurteilen. Letzten Endes leidet darunter auch die Transparenz des Begutachtungsprozesses, da die Frage, ob ein bestimmter Bestand unbedingt auf der Basis einer Editionsvorstufe bearbeitet werden muß, offensichtlich erheblich von der Lehrmeinung eines einzelnen Gutachters abhängen kann. Für die weitere Förderung sollten also klare Kategorien „angemessener Erschließung“ definiert werden, bei der dann klare Aussagen möglich sind, ob der getriebene Aufwand grundsätzlich angemessen war und ob das Ergebnis mit einem angemessenen Arbeitsvolumen erzielt wurde.“ [AlMa05, S.29]

Man kann sich nun von diesen förderpolitischen Festlegungen und von dem wissenschaftspolitischen Kompromißergebnis ein deutlicheres Bild über die Algorithmen der Informationsproduzenten machen, die sich in unserem Fall auf die Träger der DFG-Projekte und deren vertraglich gebundenen Personen und Dienstleister beziehen. Eine synchrone Beteiligung einer irgendwie geformten externen Wissenschaftsgemeinschaft bei der initialen, iterativen oder inkrementellen Erstellung und Validierung der Metadaten – wie in dem o.g. EVIA-Projekt – ist nicht konzipiert.

Fazit 2 • Damit kann der deutsche Modellkompromiß nach seinen invarianten Bestimmungen wie folgt ausgedrückt werden. Der Digitalisierungsdurchsatz pro Zeiteinheit und Euro wird hoch gehalten, in dem
a) eine explizite Einschränkung der Granularität und Sprache der Objektbeschreibungen,
b) eine geschlossene und nur organisationsinterne Schnittstelle für das Erstellen der Metadaten und
c) eine Sicherstellung der maschinellen Wiederauffindbarkeit der Objekte verfolgt wird.

2.3 Der Uniform Resource Name (URN)

Um sich nun einer impliziten Haltelinie in dem Regelungsbereich der maschinellen Wiederauffindbarkeit der Objekte anzunähern, soll zuletzt noch die eindeutige Objekt-identifizierung als wichtiger Baustein dieses projektexternen bzw. projektübergreifenden Regelungsbereichs in Bezug auf seine pragmatische Plazierung betrachtet werden. Die URN-Bezeichnertechnik ist hier als Projektgegenstand der Deutschen Nationalbibliothek (DNB) zu nennen. Es liegt nahe, den bisher genannten Regelungsinstanzen DFG-Studie und DFG-Praxisregeln in diesem neuen Zusammenhang die DNB-Regelungen beizustellen und die verteilten Aussagen aus den jeweiligen Regelungen und Empfehlungen zur URN[8] in Beziehung zu setzen.

„Uniform Resource Name [...] ist ein Uniform Resource Identifier (URI) mit dem Schema urn, der als dauerhafter, ortsunabhängiger Bezeichner für eine Ressource dient.“ [00q] Dem Bezeichner wird eine aktuelle, valide URL zugeordnet, die bei der URN-Abfrage ausgeliefert wird und den Zugriff auf die Ressource ermöglicht.

Das Konzept der URN verfolgt globale einmalige und persistente Bezeichnervergaben auf digitale Objekte. Diese unter der Oberklasse URI [00r] geführte Identifierklasse URN wird bei einer konventionellen Registrierungsinstanz pro urheberrechtsgesetzlichen

[8] Die Akronyme URN bzw. URI (beide nicht im Duden online) werden in der Fachsprache oft als feminines Kurzwort verwendet. Analog zur Abkürzung URL, die im Duden (http://www.duden.de/suchen/dudenonline/URL) den maskulinen oder femininen Genus annehmen kann, soll hier stets für URN und URI die feminine Form gewählt sein. Auch die DNB schreibt von „eine URN“ in http://www.dnb.de/DE/Netzpublikationen/URNService/urnservice_node.html (Sachstand vom 14.04.2014)

Geltungsbereich hinterlegt. Diese Geltungsbereich deckt sich in der Regel mit dem nationalen Rechtsraum. Bei der URN handelt es sich um ein syntaktisches Konstruktionsmodell für Identifikatoren nach RFC 2141 [97] in dem das Wissen um die Identität der Ressource in Klassen und Beziehungen repräsentiert wird.

Abgeleitet von der Darstellung der Aufbauregeln in RFC 2141 mit der formalen Metasprache Backus-Naur-Form, kann die Notationsschablone URN:NID:SNID-NISS auch so lesen gelesen werden, daß jedes Kolon und der Bindestrich in diesem Schema als „isSuperClassOf“ verstanden wird. Somit ist der String „URN“ als oberste Klasse definiert und der String „NISS“ als niedrigste Unterklasse.

Bei der hier behandelten URN wird die Namensraum-ID (NID) per international registrierter Konvention [00s] auf eine Klasse „National Bibliography Number (NBN)" übertragen, die wiederum aus der Bezeichnung „NBN“ selbst und der Unterklasse in einer Kodierung der national zuständigen Begründer des Identifikators z.B. „DE“ für Deutschland besteht. Als nächste Unterklasse folgt dann die Institutionenklasse „SNID“ und letztlich der eigentliche Identifier NISS (Namespace Specific String). Die Instanzen aller Klassen liegen als aneinandergereihte Literale vor, formieren damit Einzigartigkeit und werden mit einer angefügten Prüfziffer [00t] auf das Gesamtliteral versehen.

Eine URN, zusammen mit der Prüfziffer, hat folgende Syntax:
urn:nbn:de:0111-bbf-spo-10506802

2.3.1 URN und Evaluation 2005

Der Begriff „URN“ wird von den Autoren der o.g. Studie nur einmal im Studientext und zwar im Ergebnisse-Kapitel „Effektivität der Förderung“ genannt. „Auf allgemeinerer Ebene ist schwer zu verstehen, daß einerseits im Rahmen DFG-geförderter Projekte die Praktikabilität von URNs untersucht wird, während andererseits kein Mechanismus existiert, der bei den Retrodigitalisierungsprojekten die Einhaltung minimaler relevanter technischer Standards sicherstellt – was dazu führt, daß ein Angebot einer digitalen Ressource innerhalb des WWW unter ihrer numerischen IP Adresse in Linklisten eingestellt wird; aus der Sicht jedes vorstellbaren Persistenzkonzepts wohl der GAU schlechthin.“ [AlMa05]

Die Autoren nennen die URN lediglich als Teil einer Polarität mit dem anderen Pol in Form einer numerischen IP-Adresse und führen dieses Gegensatzpaar selbst als einen Indikator für die damals sehr gestreute Komplexitätsverteilung in Bezug auf viele Projektfacetten auf. Das Konzept der universal eindeutigen Identitätszuweisung auf Ressourcen im Hinblick auf ihre Entwicklung und Verwendung wird in dieser Studie nicht behandelt.

2.3.2 URN und DFG-Praxisregeln

In den o.g. DFG-Praxisregeln wird im Zusammenhang der Dublettenprüfung gefordert: „Bei Materialien der Erscheinungsjahre von 1501 bis 1800 sind das VD16, VD17 und VD18 als Referenzinstrumente für Prüfungen auf Doppeldigitalisierungen heranzuziehen. URN und PURL der Digitalisate sind an diese Verzeichnisse zu melden.“ [00g , S.6]

Hier sind zwei verschiedene Modelle von persistenten Identifikatoren synonym genannt. Über die Verschiedenheit der Modelle wird nichts berichtet. Ebenso wird später unter den Bereitstellungsregelungen diese Egalisierung wiederholt:

„Sichergestellt werden muß die Erreichbarkeit und Zitierbarkeit eines Werkes als Ganzem und die Erreichbarkeit und Zitierbarkeit von einzelnen physischen Seiten von diesem Werk. Einrichtungen sollten durch geeignete Mechanismen (PURL, URN, DOI, Handle, etc.) die Persistenz einer Ressource und der Verknüpfung zu ihr gewährleisten, um zuverlässiges Arbeiten mit den bereitgestellten Quellen in wissenschaftlichen Kontexten zu ermöglichen.
Die Erzeugung von URNs über die Deutsche Nationalbibliothek mindestens auf Werkebene wird nachdrücklich empfohlen.“ [00g , S. 39/40]

Der Abschnitt endet mit einen Verweis auf eine Fußnote, die aus einen Zeiger auf die DNB-Seite „www.persistent-identifier.de“ besteht. Nur durch eigene Recherche kann man auf dieser Website über Homepage → Einführung → Systembeispiele im Abschnitt „PURL“ lesen: „PURLs sind keine Persistent-Identifier, können jedoch in bestehende Standards wie URN überführt werden. Technisch betrachtet wird bei PURL der existierende Internet-Standard ‚HTTP-redirect‘ angewendet, um PURLs in die URLs aufzulösen.“ [00u]

Diese Menge der von der DFG angebotenen Wahlmöglichkeiten wird weiter unten wiederholt und im zweiten, anschließenden Satz jedoch wieder kontextbezogen aufgelöst: „Einrichtungen sollten durch geeignete Mechanismen (PURL, URN, DOI, Handle, etc.) die Persistenz einer Ressource und der Verknüpfung zu ihr gewährleisten, um zuverlässiges Arbeiten mit den bereitgestellten Quellen in wissenschaftlichen Kontexten zu ermöglichen. Die Erzeugung von URNs über die Deutsche Nationalbibliothek mindestens auf Werkebene wird nachdrücklich empfohlen.“ [00g , S. 50]

2.3.3 URN und DNB

Die Deutsche Nationalbibliothek führt diesen URN-Standard im Rahmen des Zugangs zu digitalen Objekten mit folgender Beschreibung ein: „Persistent Identifier sind eindeutige, standortunabhängige Identifikatoren für digitale Objekte, um über lange Zeiträume und eventuelle Systemwechsel hinweg einen zuverlässigen Zugriff auf diese Ressourcen gewährleisten zu können.“ [00v]

Die DNB bietet für die Unterstützung der Ablieferung von sammelpflichtigen Netzpublikationen an sie selbst die Vergabe von urn:nbn:de als Persistent Identifier an. Für die Partner des Sammelpflichtverhältnisses und für sonstige Interessierte ist eine „Policy für die Vergabe von URNs im Namensraum urn:nbn:de“ maßgebend. [12] „Die Deutsche Nationalbibliothek vergibt und verwaltet diese URNs aus dem Namensraum „urn:nbn:de“ und bietet einen URN-Resolving-Dienst für Deutschland und die Schweiz an.“ [00w]

Exakter als in den DFG-Praxisregeln bezeichnet die DNB in ihrer o.g. Policy: „Uniform Resource Names (URN) sind Persistent Identifier (wie z.B. auch DOI, ARK oder Handle) und dienen damit der dauerhaften und ortsunabhängigen Identifizierung von Ressourcen.“ [12] Die Policy bleibt damit bei dem Modell der Objektidentifizierung. Die in den DFG-Praxisregeln genannte PURL drückt hingegen das Modell einer eindeutigen Adressidentifizierung aus und formuliert damit einen anderen Bildbereich.[9]

Die DNB verfolgt mit diesen Regelungen

[9] Der Terminus Bildbereich bzw. Bild oder Bildmenge ist als mathematische Funktion zu verstehen. Streng genommen können unter einer identifizierten Adresse mehrere Objekte „wohnen“. Ein identifiziertes Objekt ist eindeutiger.

- eine verpflichtete Teilnahme an der Langzeitarchivierung
- eine Beschränkung der Teilmenge Institutionenklasse (SNID) aus der Gesamtmenge aller publizierenden juristischen und natürlichen Personen des Rechtsraums aus organisatorischen Gründen
- ein verläßliches Gateway, das die registrierten URN-URL-Paare als aktive Ressourcenanzeiger immerwährend brauchbar macht
- eine eindeutige URN-Objektbeziehung, wobei ein Objekt nur eine einzige URN besitzen darf, auch wenn das Objekt bei mehreren Institutionen als Kopie geführt wird.

Kritik 2

Es läßt sich abschließend nach der Gegenüberstellung aller drei Regelungsinstanzen feststellen, daß die DFG-Studie die Objektidentifizierung nicht behandelt und die DFG-Praxisregeln keine eindeutige Stellung zu dem Modell der Objektidentifizierung einnehmen. PURL und URN werden in fachlich nicht korrekter Weise in den DFG-Praxisregeln einander gleichgestellt. Für das PURL-Konzept ist überdies bei der DNB keine zentrale Registratur nach der Art der URN vorhanden. Die PURLs verbleiben jeweils auf der Referenzebene der einzelnen Gateways, die nach dem Domain Name System entsprechend zuständig sind.

2.3.4 Objekt und URN

Im Anhang 2 mit dem Titel „Vergabe von URNs im Namensraum urn:nbn:de für verschiedene Objektarten“ drückt nachfolgende Aussage die relative Entscheidungsfreiheit aus, welche „inhaltlich geschlossene Einheit“ der Netzpublikation bei der DNB per URN registriert wird: „Entscheidend ist stets, daß die urn:nbn:de das Archivobjekt identifiziert.“ [12 , S. 13] Der Begriff „Archivobjekt“ wird nur bei dem Strukturelement Zeitschriftenheft und Zeitschriftenartikel genannt. Dagegen wird an vielen Stellen in der Policy der allgemeinere Terminus „Objekt“ benutzt. Der „URN-Service (sieht sich) heute mit vielen verschiedenartigen Objektarten konfrontiert.“ (Siehe Haupttext, [12, S. 4])

Als weitere Objektart ist die Klassenbezeichnung „Digitalisate“ zu nennen. In der Beschreibung dieser Objektart erkennt man die formatbezogene Verbindung zum METS-

Standard wieder. „Die Einzelseiten liegen als einzelne Bilddateien in einer Paketstruktur vor. Die Bilddateien sind zwar meist prinzipiell einzeln adressierbar, haben aber eine gemeinsame Basisadresse und werden vom Archiv als Einheit behandelt.“ [12, S. 14] Die Policy empfiehlt „die Vergabe der urn:nbn:de auf Werkebene mit zusätzlicher Möglichkeit der Verwendung von Fragmentadressierungen für Teilobjekte.“ [12, S. 14]

Diese 'Verschiedenartigkeit der Objektarten' motiviert wiederum eine Tendenz des symbolischen Systems bzw. des Modells von analogen Originalen, die folgendermaßen beschrieben werden kann: Das Konzept tendiert bei diesem Modelleinsatz mehr in Richtung Strukturanalogie und nicht in die Richtung einer Funktionsanalogie. Die Policy „richtet sich an Institutionen, die ihre digitalen Publikationen persistent identifizieren, um eine zuverlässige Zitierbarkeit zu gewährleisten.“ [12, Haupttext S. 4] Mit einer Analogie soll das Modell eine bestimmte Aufgabe lösen, weil das Original nicht direkt zur Verfügung steht.

Kritik 3

Die DNB bemüht sich mit der veröffentlichen Policy (neben den vielfältigen Onlineinformationen, insbesondere bei www.persistent-identifier.de) „die Transparenz bei der Verwendung einer urn:nbn:de zu verbessern.“ [12, Haupttext S. 4] Das scheint jedoch nicht optimal gelungen zu sein.

a) Es sind die Intentionen zur (Langzeit-)Archivierung und zur Netzpublikation als Basis für das Modellieren von Forschungsinfrastruktur nicht deutlich genug im gegenseitigen Verhältnis explizit gemacht - unabhängig davon, wer welche Absicht verfolgt. Sodann sind die Sprünge in der Verwendung der Termini - wie o.g. Archivobjekt, Netzpublikation, Objekt, Digitalisat, Teilobjekt, digitale Publikation – nicht nachvollziehbar.

b) Der Adressierstatus Archiveinheit wirkt in der logischen Dimension seiner Bedeutung als Grenzschicht für tiefer liegende Strukturen. Egal wo diese Grenze gezogen wird: die tiefer liegenden Strukturen wie z.B. Einzelseiten sind damit für eine Adressierung als Diskursobjekt aus dem URN-Konzept hinausgedrängt.

c) Außerdem ist die (b) heilende „Fragmentadressierung für Teilobjekte“[10] nicht klar defi-

[10] Unklar ist, ob es sich bei dieser Fragmentadressierung um das Modell von Tim Berners Lee in RFC 1630 [Bern94] handelt, deren Trennzeichen das "#" ist. „The hash ("#", ASCII 23 hex) character is reserved as

niert. Es ist aus der Lektüre des Textes nicht entscheidbar, ob diese Fragmente zur URN gehören oder auf Anbieterseite - womöglich nicht persistent - im Zugriffssystem angefügt werden. Die Fragmentadressierung wird zumindest nötig, wenn man in den nicht URN-referenzierten Teilstrukturen des URN-referenzierten Objekts eine wiederholbare und nachvollziehbare Exploration betreiben möchte.

Wenn man nun aus den Betrachtungen der DFG-Richtlinien, der DFG-Viewer-Passierstelle und der URI-Umkreisungen eine Schlußformel ableiten wollte, so findet sich im PDF-Dokument [Funk09], und in der XML-Version [EnFu00a] zum „zvdd/DFG-Viewer METS-Profil – Version 2.0" dafür einen Werkzeugbegriff, der seines Gleichen sucht: der Page-Turner. Dieser Begriff betitelt in der offiziellen Dokumentation des METS-Standards bei der Library of Congress [00x] eine Anwendungsdomäne in deren Beispielsammlung[11].

Es geht also ums Blättern - nach wie vor. So stark noch das Blättern in einer elektronischen Modellinstanz von einem einmaligen physischen Objekt mit bestimmter Lokalisierung in Raum und Zeit eine Anziehungskraft ausübt, so fest bindet das METS-Datenformat mit seiner hierarchischen XML-Syntaxleistung sein Domänenkonzept an ineffiziente Austauschwege mit der übrigen Welt. Beim dem Goobi-METS mit seinem Page-Turner-Modell mangelt es an einer Öffnung von taxonomischen Verbindungen zwischen Strukturelementen in die Richtung zu vernetzbaren Informationen vielfältiger Art mit logischen Relationen in expliziter Form. So bildet bei dem Leistungsmerkmal der Einzelseiten-Repräsentation die Referenzierbarkeit einen kontrollillusorischen Qualifikationsüberschuß.

---

a delimiter to separate the URI of an object from a fragment identifier." Die oben erwähnte RFC 2141[97] notiert dazu: „RFC 1630 [...] reserves the characters "/", "?", and "#" for particular purposes. The URN-WG [WG = working group, der Verf.] has not yet debated the applicability and precise semantics of those purposes as applied to URNs. Therefore, these characters are RESERVED for future developments. Namespace developers SHOULD NOT use these characters in unencoded form, but rather use the appropriate %-encoding for each character." [97 , S, 3]

[11] Die Anwendungsdomänen sind: Bibliographic Record, Page Turners, Internationalization, Maps & Geographic, Image with Text, PDF and Other Document Types, Pictorial Material, Image with Video, Multiple Pictorial Images, MrSid and Various Other Static Image Formats, Sheet Music, Sound Recording , Realia, Compact Disc , Musical Score and Parts, Serials, Video with Transcript.

3. Goobi als Softwarepaket für die Digitalisierung

3.1 Einleitung

„Goobi ermöglicht Digitalisierungsprojekte in großen und kleinen Bibliotheken, Archiven, Museen und Dokumentationszentren. [...] Goobi ist ein Softwarepaket für die Digitalisierung“[12] Das Paket beinhaltet Softwaremodule, deren wichtigste, von einander getrennten Anwendungen „Goobi.Production“ [00c] und „Goobi.Presentation“ [00d] sind. „Goobi wurde zunächst in den ersten vier Jahren (2004-2008) federführend von der Staats- und Universitätsbibliothek Göttingen im Rahmen DFG-geförderter internationaler Digitalisierungsprojekte entwickelt.“ [00y] Im September 2012 hat sich der Verein „Goobi. Digitalisieren im Verein“ gegründet, der frühere Auseinanderentwicklungen potenter Akteure auf dem Gebiet der Quellcodeerstellungen und -verwertungen für Goobi zu einer mehr kooperierenden Gemeinschaft von Anwendern, Softwareentwicklern und Dienstleistungsunternehmen in Digitalisierungsprojekten motiviert. Der neue Organisationsrahmen erlaubt damit das aus der Sache herrührende notwendige, aufwendige und teure Software-Release-Management von Goobi-Modulen zu finanzieren. Den überwiegend öffentlich-rechtlich verfaßten Gemeinschaftsmitgliedern wird über Mitgliedschaftsbeiträge eine haushaltsrechtlich machbare Unterstützung der ideellen Vereinsziele des Goobi-Projektes ermöglicht.

Das Wissen und die Technik zur Erstellung und zum Lesen von METS-Dokumenten, die schließlich im o.g. DFG-Viewer angezeigt werden können, ist in Regelwerken und Programmfunktionen verteilt abgelegt. Das seitenbasierte Dokumentenmodell ist durch die Repräsentation einer logischen Struktur und einer physischen Struktur und deren gegenseitigen logischen Verbindungen beschrieben.

[12] Aus der Portalseite von http://www.goobi.org/ - zuletzt abgerufen: 14.04.2014

Es ist zu vermuten, daß der Verein auch im Bereich der Wissensvermittlung neue Wege beschreiten wird. Die mit Goobi befaßten Personengruppen sind aus eigener Erfahrung durch den Besuch einiger Goobi-Anwendertreffen wie folgt typischerweise einzuteilen:

- Großeinrichtungen mit vorhandenem Personal, das an der Quellcodeentwicklung mitarbeitet
- Erfahrene Goobi-Anwender mit den personellen Kapazitäten zur selbständigen Implementierung der Software in einen Digitalisierungsworkflow
- Wenig erfahrene Goobi-Anwender, bisweilen ohne ausreichende Personalkapazität und deswegen auf assistierte Implementierung der Software und ggf. fortlaufende Betreuung durch kommerzielle Dienstleister zurückgreifend.

Für die beiden ersten Gruppenmitgliedschaften sind vorhandene Schlüsselqualifikationen in schlagwortartig benannten Bereichen Java-Programmierung, Linux-Betriebssysteme, Apache, Tomcat, Lucene-Index, Typo3-CMS sowie Typoskript in unterschiedlichem, eher größeren Ausmaß Voraussetzung. Mittlerweile wurde eine weitere Kommunikationsplattform für Experten neben den allgemeinen Anwendertreffen eingerichtet. Für die letzte Gruppe der obigen Auflistung stehen die Informationsseiten der mittlerweile dem Verein rechtlich zugehörigen Website www.goobi.org zur Verfügung und ein seit August 2007 bestehendes Wiki.[13]

Zwischen dem Informationsniveau der Quellcodes samt den darin befindlichen Kommentaren und den Informationsniveau der Seiten aus der Goobi.org-Website und dem Wiki besteht eine Vermittlungslücke. Für Unerfahrene in den Schlüsseltechnologien bieten die Wiki- und Webseiten keine ausreichenden, didaktisch gestalteten Explorationswege, die ein Messestandniveau vertiefen könnten. Für die Pflege dieser Informationsschicht hat sich noch keine verfaßte Redaktion gebildet. Die früheren im Wiki engagierten Personen sind mit Arbeiten am Quellcode in das englischsprachig geführte Software-Release-Management erschöpfend beschäftigt umgezogen. Es ist für interessierte Anwender mühselig, die als Quellen angegebenen offiziellen Dokumente zu METS, MODS, UGH - Java Metadaten Bibliothek, METS-Profil des DFG-Viewers,

[13] http://wiki.goobi.org/index.php/Hauptseite - zuletzt abgerufen: 14.04.2014

Goobi-Regelwerk-Vorlagen zu einer kompakten Orientierungsgrundlage zusammenzufassen.

Die nachfolgenden Abschnitte sollen (im eingeschränkten Kontext des Untersuchungsgegenstands und -auftrags) jedoch eine neue, bei weitem nicht umfassende Entfaltung der Sachverhalte versuchen.

3.2 Das generische Dokumentenmodell für Goobi

Die „UGH - Java Metadaten Bibliothek“ [EnFu00b] begründet einen generischen Unterbau für das Modellieren beliebiger Dokumentklassen. Das angestrebte universelle Dokumentenmodell ist „von unterliegenden Formaten zur Beschreibung von Metainformationen unabhängig“. [EnFu00b , S. 4]

Es existiert - neben anderen Klassen - eine Klasse zur Serialisierung eines Dokuments im METS/MODS-Datenformat, das Gegenstand dieser Untersuchung ist. In diesem Dokumentenmodell sind Strukturdaten beschreibbar. „Hierzu wird allerdings ein recht weitgehender Strukturbegriff verwendet, der nicht nur den internen, hierarchischen Aufbau eines Dokumentes betrifft, sondern das Dokument an sich ebenfalls als Teil dieser Struktur versteht- als bibliographische Einheit 'Dokument'.“ [EnFu00b, S. 5] Das Dokument ist also selbst eine Klasseninstanz der Klasse Strukturelement.

Zur Veranschaulichung dienen die beigefügten Diagramme. Die für diese Untersuchung neu erstellten Diagramme haben als Orientierungshilfe eine Spalten- und Reihennotation an den Blatträndern. Orientierungshinweise werden im nachfolgenden Text geschrieben im Sinne von „Diagramm 1 Spalte E bis F Reihe 5 bis 6,"und dies in Kurzform: (Dg1 EF5-6).

Diagramm 1 bildet die Struktur der internen XML-Datei „meta.xml“ ab. Es handelt sich hier nicht um ein valides METS-Format, wie es später dem Viewer über einen Export aus „Goobi.production“ bereit gestellt wird. Außerdem sind in diesem Diagramm nicht alle Attribute und Werte abgebildet. Die Darstellungsauswahl soll eine syntaktische Orientierung über die Objektbeziehungen geben. Dies gilt auch für die anderen Diagramme.

„Neben dieser logischen Struktur [(Dg1 DE6)] kann das Modell die physische Struktur [(Dg1 FG7)] als zweite Struktur abbilden. Diese Struktur speichert üblicherweise die 'gebundene Einheit' [(Dg1 G7) = PHYS_0000 und (Dg1 CD6) = LOG_0000] sowie ihre Untereinheiten [(Dg1 C7-8) = LOG_0005 etc. und (Dg1 BH10) = PHYS_0024ff.] (d.h. die einzelnen Seiten des Werkes).“ [EnFu00b , S. 5]

In Goobi.Production ist mit der logischen Struktur die Perspektive des Dokuments als Band eingeführt: als (gebundener) Jahrgangsband einer Zeitschrift, als einzelner Band einer Monographie oder als Einzelband eines mehrbändigen Werkes. Die vom Programm behandelbaren Objekte werden als Goobiterminus „Vorgänge“ genannt. Mit „Goobi.Production“ ist das Erstellen der digitalen Publikationen als mehrschrittiger Workflow gestaltet. Das digitalisierte Werk ist das Ergebnis eines Vorgangs. Eine abstrakte Dachstruktur wird zudem als Anker zum übergeordneten Werk, wie die Zeitschrift, das Sammelwerk oder das mehrteilige Einzelwerk behandelt. Dieser Anker „besitzt keine Inhaltsdateien - also kein Image Set mit Bild-Dateien, er besteht lediglich aus einer Struktureinheit, die über deskriptive Metadaten verfügt.“ [EnFu00b , S. 6] Der Anker verweist auf das Metadatenset und den Identifier einer externen Datenbank z.B. von Bibliotheksverbünden, Lokaldatenbank oder der Zeitschriftendatenbank (ZDB). Über dieses Dachobjekt können in einer Struktursicht die Teilwerke z.B. durch eine Listendarstellung präsentiert werden. Umgekehrt ist festzustellen, daß die Ankerhierarchie nur eine Schicht mit einer Instanz aus der Klasse der Anker haben kann Es besteht keine Möglichkeit weitere hierarchisch höhere oder parallele Ankerinstanzen mit der Instanz des Dokumentenbandes bzw. des damit übergeordneten Objekts zu verknüpfen.

3.2.1 Hierarchie

„Sowohl in der logischen als auch in der physischen Struktur können Einheiten beliebig tief hierarchisch geschachtelt werden, wobei jede Einheit nur eine einzige Elterneinheit besitzen kann.“ [EnFu00b , S. 5] Aus der Aussage folgt, daß ein gerichteter zyklenfreier Graph, also ein Baum die Ausprägung der jeweiligen Hierarchie ist.

```
<mets:structMap TYPE="LOGICAL">
- <mets:div DMDID="DMDLOG_0000" ID="LOG_0000" TYPE="Monograph">
  - <mets:div DMDID="DMDLOG_0001" ID="LOG_0001" TYPE="Chapter">
    - <mets:div DMDID="DMDLOG_0002" ID="LOG_0002" TYPE="Poem">
        <mets:div DMDID="DMDLOG_0003" ID="LOG_0003" TYPE="Remarks" />
      </mets:div>
    </mets:div>
    <mets:div DMDID="DMDLOG_0004" ID="LOG_0004" TYPE="Chapter" />
    <mets:div DMDID="DMDLOG_0005" ID="LOG_0005" TYPE="Chapter" />
    <mets:div DMDID="DMDLOG_0006" ID="LOG_0006" TYPE="Chapter" />
    <mets:div DMDID="DMDLOG_0007" ID="LOG_0007" TYPE="Chapter" />
    <mets:div DMDID="DMDLOG_0008" ID="LOG_0008" TYPE="Chapter" />
  </mets:div>
</mets:structMap>
```

Abb. 1: StructMapLogTop

Bei dem seitenbasierten Dokumentenmodell ist nach bisheriger Erfahrung eine tiefere Schachtelung in der physischen Struktur nicht motivierend. „Innerhalb des physischen <structMap> Elements wird die Seitenstruktur durch Elemente wiedergegeben, die einem obersten Element untergeordnet sind. Dieses oberste Element umfaßt die Seiten, die die bibliographische Einheit repräsentieren. Daher muß dessen TYPE Attribut immer den Wert 'physSequence' besitzen. " [EnFu00a] (siehe Abb. 2)

```
<mets:structMap TYPE="PHYSICAL">
  - <mets:div TYPE="physSequence" ID="PHYS_0000" DMDID="DMDPHYS_0000">
    - <mets:div TYPE="page" ID="PHYS_0001" CONTENTIDS="urn:nbn:de:0111-bbf-spo-8122706" ORDER="1" ORDERLABEL="147">
        <mets:fptr FILEID="FILE_0000_PRESENTATION"/>
        <mets:fptr FILEID="FILE_0000_DEFAULT"/>
        <mets:fptr FILEID="FILE_0000_MIN"/>
        <mets:fptr FILEID="FILE_0000_MAX"/>
        <mets:fptr FILEID="FILE_0000_THUMBS"/>
      </mets:div>
```

Abb. 2: StructMapPhysSequence

3.2.2 Typisierung

Die Abb. 1 StructMapLogTop zeigt eine Hierarchie, deren typisierten Elemente zum Beispiel *monograph* ⊃ *chapter* ⊃ *poem* ⊃ *remarks* repräsentieren, und dies in einer entsprechenden Schachtelung der <mets:div>-Elemente [siehe (Dg1 CD6-8)], die einer Taxonomie von bestimmten Dokumentarten folgt.

Eine Überlappung von Hierarchien, die naturgemäß vorliegt, wenn man Strukturelementen aus unterschiedlichen Hierarchieebenen wiederholt die gleichen Seiten zuweist, wird mit Verknüpfungen zwischen voneinander getrennten Hierarchien gelöst. „Jeder Struktureinheit können ferner Inhaltsdateien (Contentfiles) zugeordnet sein, wobei

eine m:n Relation zwischen Struktureinheiten und Inhaltsdateien möglich ist.“ [EnFu00b , S. 5] (Dg1 J4-9)

„Ferner können Einheiten aus der logischen und der physischen Struktur miteinander verknüpft werden, um so bspw. das Verhältnis zwischen einem Kapitel und seinen entsprechenden Seiten wiederzugeben.“ [EnFu00b , S. 5] So ist (wie nachfolgend in Abb. 3 zu sehen) die physische Seite PHYS_0003 zugehörig zu den logischen Strukturen Band (LOG_0000), zum Heft (LOG_0001), zum Artikel (LOG_0002) und auch noch Artikel (LOG_0003), da der Artikelwechsel auf der gleichen Seite stattfindet.

```
<mets:smLink xlink:to="PHYS_0034" xlink:from="LOG_0000" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0001" xlink:from="LOG_0001" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0002" xlink:from="LOG_0001" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0003" xlink:from="LOG_0001" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0004" xlink:from="LOG_0001" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0005" xlink:from="LOG_0001" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0006" xlink:from="LOG_0001" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0007" xlink:from="LOG_0001" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0008" xlink:from="LOG_0001" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0002" xlink:from="LOG_0002" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0003" xlink:from="LOG_0002" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0003" xlink:from="LOG_0003" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0009" xlink:from="LOG_0004" xmlns:xlink="http://www.w3.org/1999/xlink" />
<mets:smLink xlink:to="PHYS_0010" xlink:from="LOG_0004" xmlns:xlink="http://www.w3.org/1999/xlink" />
```

Abb. 3: SmLink-xlink-to-from • siehe auch (Dg1 JK3-9)

„Sowohl Inhaltsdateien als auch Struktureinheiten können Metadaten besitzen. Ein Metadatum zeichnet sich dadurch aus, daß es einen bestimmten Typ sowie einen Wert besitzt und entweder einer Struktureinheit oder einer Inhaltsdatei zugeordnet ist; es ist ein Typ-Wert Paar, welches zur Beschreibung des verknüpften Objekts dient. Ein Metadatum kann immer nur einem Objekt zugeordnet sein.“ [EnFu00b , S. 5] Ein Metadatum kann also nicht selbst ein Objekt sein. Der Begriff „Typ“ ist in diesem Zusammenhang nach [Wagn00a] als eine Klasse äquivalenter Exemplare zu verstehen. Die Verwendung des Begriffs „Typ“ ist bei [EnFu00b] verwirrend, weil in XML dafür der Begriff „name“ (Abb. 4) verwendet wird.

Als Beispiel für die Anbindung eines Metadatums an ein Objekt dient hier (Abb. 4) eine Beschreibung eines logischen Strukturelements „metadata“ aus dem Namespace „goobi“ mit dem XML-Elementattribut „name“ und XML-Attributwert „TitleDocMain“ und dem Elementinhalt „Kapitel 5".

```
<mets:dmdSec ID="DMDLOG_0005">
  - <mets:mdWrap MDTYPE="MODS">
      - <mets:xmlData>
          - <mods:mods xmlns:mods="http://www.loc.gov/mods/v3">
              - <mods:extension>
                  - <goobi:goobi xmlns:goobi="http://meta.goobi.org/v1.5.1/">
                        <goobi:metadata name="TitleDocMainShort">Kapitel 5</goobi:metadata>
                        <goobi:metadata name="TitleDocMain">Kapitel 5</goobi:metadata>
                        <goobi:metadata name="_urn">urn:nbn:de:0111-bbf-spo-14205681</goobi:metadata>
                    </goobi:goobi>
                </mods:extension>
            </mods:mods>
        </mets:xmlData>
    </mets:mdWrap>
</mets:dmdSec>
```

Abb. 4: dmdSec-DMDLOG

Eine hierzu analoge „Typ/Wert“-Zuordnung ist in einem zum „Kapitel 5“ zugewiesenen physischen Strukturelement zu finden mit dem Typ "_urn“ und mit dem Literal einer URN als Wert. (Abb. 5)

```
<mets:dmdSec ID="DMDPHYS_0026">
  - <mets:mdWrap MDTYPE="MODS">
      - <mets:xmlData>
          - <mods:mods xmlns:mods="http://www.loc.gov/mods/v3">
              - <mods:extension>
                  - <goobi:goobi xmlns:goobi="http://meta.goobi.org/v1.5.1/">
                        <goobi:metadata name="_urn">urn:nbn:de:0111-bbf-spo-14205956</goobi:metadata>
                    </goobi:goobi>
                </mods:extension>
            </mods:mods>
        </mets:xmlData>
    </mets:mdWrap>
</mets:dmdSec>
```

Abb. 5: dmdSec_DMDPHYS

Es befinden sich außerdem Metadaten zum übergeordneten Werk bei Zeitschriftenbänden oder Einheiten von mehrbändigen Werken im obersten logischen Strukturelement, das bei Goobi mit <mets:dmdSec ID=„DMDLOG_0000"> identifiziert wird (Abb. 6) und als der oben benannte Anker verstanden wird.

```
<mets:dmdSec ID="DMDLOG_0000">
  - <mets:mdWrap MDTYPE="MODS">
      - <mets:xmlData>
          - <mods:mods xmlns:mods="http://www.loc.gov/mods/v3">
              - <mods:extension>
                  - <goobi:goobi xmlns:goobi="http://meta.goobi.org/v1.5.1/">
                      <goobi:metadata name="CatalogIDDigital">319141853</goobi:metadata>
                      <goobi:metadata name="CatalogIDAllegro">123456789</goobi:metadata>
                      <goobi:metadata name="CatalogIDSource">31914185a</goobi:metadata>
                    [...]
                      <goobi:metadata name="_digitalOrigin">reformatted digital</goobi:metadata>
                    - <goobi:metadata name="Author" type="person">
                        <goobi:lastName>Snell</goobi:lastName>
                        <goobi:firstName>Christian Wilhelm</goobi:firstName>
                        <goobi:identifier>151228582</goobi:identifier>
                        <goobi:displayName>Snell, Christian Wilhelm</goobi:displayName>
                      </goobi:metadata>
                  </goobi:goobi>
              </mods:extension>
          </mods:mods>
      </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
```

Abb. 6: DMDLOG_0000 • Goobi-Produktionsdatenformat (mit Ausschnitt der Metadaten im Goobi-Namespace)

Das Modell bereitet eine abstrakte logische Struktur von Elementen in hierarchischer Gestaltung vor. (Dg0 B4-9) (Abb. 7)

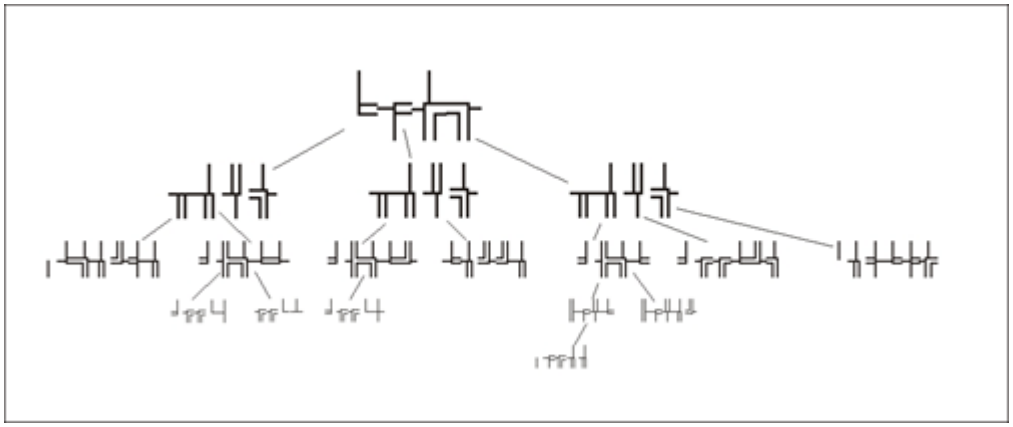


Abb. 7: Schematische Darstellung der Dokumentstruktur (kein Schriftsatzfehler!)

3.2.3 Der Regelsatz als darüber gelegter Schema-Layer

Zur Instanziierung von Strukturtypen und Metadatentypen wird ein separates XML-Dokument genutzt, in dem in einer für das Programm verarbeitbaren Syntax die Typen definiert werden. Zunächst werden die Metadatentypen definiert und damit ein Vokabular bzw. Vokabeln für eine Strukturdatentypologie angelegt.

```
<MetadataType>
      <Name>TitleDocMain</Name>
      <language name="de">Haupttitel</language>
      <language name="en">main title</language>
      <language name="es">Título principal</language>
</MetadataType>
```

(Für das mehrsprachige Benutzerinterface sind Vokabeln der eigenen Sprache erfaßbar)

Für jeden Strukturtyp wird in einer jeweils eigenen XML-Sequenz eine individuelle Taxonomie zugewiesen, in der die erlaubten und ebenso instanziierten Strukturtypen als Kindelemente aufgeführt sind. (Abb. 8)

```
<DocStrctType>
            <Name>PeriodicalIssue</Name>
            <allowedchildtype>TitlePage</allowedchildtype>
            <allowedchildtype>Article</allowedchildtype>
            <allowedchildtype>Miscella</allowedchildtype>
            <allowedchildtype>Review</allowedchildtype>
            [...]

<DocStrctType>
            <Name>Article</Name>
            <allowedchildtype>Table</allowedchildtype>
            <allowedchildtype>Illustration</allowedchildtype>
            <allowedchildtype>Poem</allowedchildtype>
            <allowedchildtype>Chapter</allowedchildtype>
            [...]
```

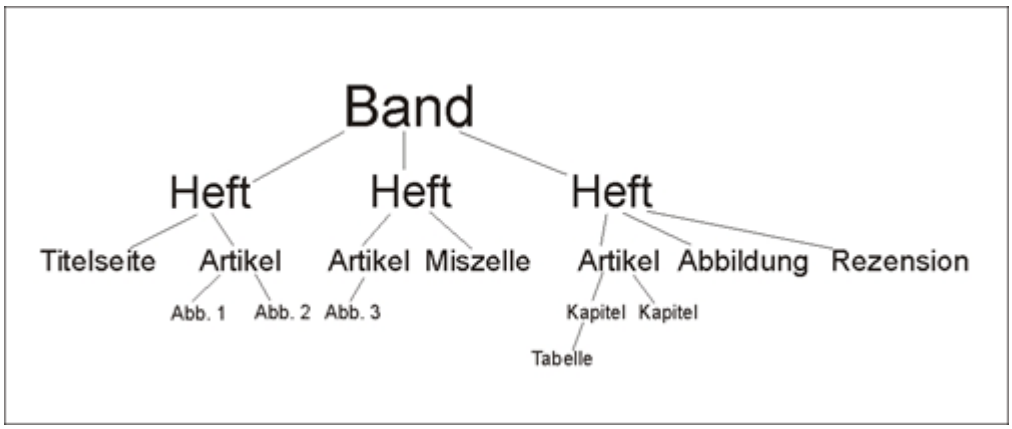


Abb. 8: Schematische Darstellung von instanziierten Dokumentstrukturtypen

In derselben XML-Sequenz zur Strukturtypinstanziierung sind die Metadatentypen mit optionaler Angabe einer Kardinalität über das Elementattribut „num“ aufgelistet.

```
[...]
<DocStrctType>
<Name>Article</Name>
<metadata num="+" DefaultDisplay=„true“>TitleDocMain</metadata>
<metadata num="*" DefaultDisplay="true“>TitleDocSub1</metadata>
<metadata num="*" DefaultDisplay="true“>TitleDocMainShort</metadata>
<metadata num="*">TitleDocParallel</metadata>
<metadata num="*">VariantTitle</metadata>
[...]
```

Die Kardinalitätskodierung in Goobi-eigener Konvention ist wie in nachfolgender Abb. 9 zu lesen. Ist kein Attribut angegeben, wird standardmäßig der Wert "*" angenommen.

„*“	kein mal oder beliebig oft (0...n)
„+“	ein mal oder beliebig oft (1...n)
„1o“	kein mal oder genau einmal (0...1)
„1m“	genau einmal (1)

Abb. 9: Kardinalitätskodierungen im Regelsatz von Goobi

Eine tiefere Exploration der UGH - Java Metadaten Bibliothek kann hier aus Gründen des Umfangs der Arbeit und der selektiven Motivation nicht verfolgt werden. Im „Ausblick“-Kapitel des Dokuments zur UGH - Java Metadaten Bibliothek wird jedoch ein Desiderat angesprochen, das nicht nur die Handhabbarkeit des Software-Dokumentpflege-Verbundes verbessern könnte, wie es dort formuliert wird:

„Weiterhin kann der Anwendungszweck der UGH Bibliothek ausgedehnt werden, wenn diese mittels einer Persistenzschicht die einzelnen Objekte direkt ansprechen könnte, ohne über Serialisierungsklassen zu gehen. Dieser Verzicht der Serialisierungsschicht würde dazu führen, daß zum Ändern/Löschen/Aktualisieren von einzelnen Objekten nicht immer das komplette Dokument gelesen und geschrieben werden müsste. Serialisierungsklassen würden dann lediglich zum Import und Export genutzt werden. Auf eine solche Klassenbibliothek ließe sich auch direkt ein Repository aufbauen, welches aufgrund der identischen API all jene Tools nutzen kann, die derzeit auf beispielsweise METS-Dateien anwendbar sind.“ [EnFu00b , S. 38]

Bei der Handhabbarkeit zeigt es sich, daß massenhafte Änderungen an den Metadaten zu zeitraubenden Lade- und Schreibvorgängen führen und zu einem Arbeitsbelastungsproblem werden. Ebenso wie in Bibliotheksdatenbanken ständig

Korrekturen und Ergänzungen spontan vorgenommen werden, sind die Goobi-Metadaten im Rahmen des Projektes von Scripta Paedagogica Online (SPO) von solchen wiederholten Korrekturvorgängen ausdrücklich nicht ausgeschlossen.

In diesem soeben zitierten Ausblick [EnFu00b , S. 38] steckt außerdem der Kern für eine paradigmatische Entscheidungsfrage, die da ist, ob man nicht weitere Objekte als eindeutig zu identifizierbare, formal-semantisch beschriebene Entitäten interpretierbar machen sollte. Außerdem wäre die Überlegung angebracht, ein terminologisches Wissen in einer anderen Form zu notieren und soweit explizit zu machen, so daß automatische Schlußfolgerungen möglich sind. In einem Regelsatz von Scripta Paedagogica Online (SPO) zur Beschreibung von Zeitschriftenaufsätzen[14] sind beispielsweise notiert (jedoch nicht alle genutzt):

108	Metadatentypen
80	Strukturdatentypen
1.330	„allowedchildtype“-Zuweisungen bei den Strukturdatentypen
1.137	Zuweisungen von Metadatentypen zu den Strukturtypen

Tabelle 1: Statistik zum Regelsatz allegro.xml

Dieser Regelsatz muß manuell mit einem Texteditor gepflegt werden. In den 4.570 Zeilen Anweisungen, die in diesem XML-Dokument im Top Level-Element <preferences> eingetragen sind, werden 1.330 Aussagen zu erlaubten 80 möglichen Kindelementen (siehe Tabelle 1) redundant auf die 80 Strukturdatentypen verteilt. Bei insgesamt 1.137 Metadatenfeldzuweisungen zu den 80 Strukturtypen wird beispielsweise 70 mal „TitleDocMain“ (Hauptsachtitel bzw. Titel von Bänden) zugewiesen. Die Pflege der Regelsätze erweist sich als risikobehaftet, weil es keine Prüfwerkzeuge gibt, die vor dem (Wieder-)Einsatz des Regelwerks in der Produktionsumgebung eine Konsistenzprüfung bzw. Validierung durchführen. Der durchaus XML-wohlgeformte Regelsatz kann erst zu fortgeschrittener Zeit einen Laufzeitfehler auslösen, wenn endlich das betroffene Metadatum in einem Goobi-Vorgang involviert ist, der im Zuge von Veränderungen am Datenset neu geschrieben wird. So muß nach bisheriger Erfahrung die Regelsatzdatei im

[14] In der Sammlungsübersicht von SPO ist das die Kollektion „Pädagogische Zeitschriften“ http://goobiweb.bbf.dipf.de/viewer/browse.xhtml

Nachzug manuell auf Schreibfehler bzw. Denkfehler hin untersucht werden. Dies erfolgt unter dem erschwerenden Umstand, daß die Fehlermeldungen des Systems in manchen Fällen keinen deutlichen Hinweis geben, wo das Problem liegt.

3.3 Zwei Goobi-Welten: PURL und URN

Unterschiedliche pragmatische Anforderungen an die Referenzierbarkeit haben in den letzten Jahren unterschiedliche Entwicklungszweige des Quellcodes der Goobi-Produktionsumgebung entstehen lassen. Damit hat sich die Verwendung von PURLs (z.B. Digizeitschriften.de[15]) für die digitalen Publikationen und die Verwendung von URNs (z.B. SLUB Dresden[16] , SPO[17]) heterogen vollzogen. Plug-In-Mechanismen der Firma intranda GmbH[18] (Göttingen) erweiterten „Goobi.Production“ um die Möglichkeit, in die Metadaten der digitalen Publikationen auf der Seitenebene URNs aufzunehmen und zu veröffentlichen.

Wie bei dem Goobi-Produktionsmodul, vollzog sich eine Teilung der Softwareentwicklung auf der sog. Viewerseite mit „Goobi.Presentation“ (open source, Typo3-Basis) [00d] und dem „intranda viewer“ (proprietär, Java-Basis). [00e] Der „intranda viewer“ initiierte eine OAI-PMH-Schnittstelle und konnte damit die URN-Registrierung der DNB über ein Harvestingverfahren ermöglichen. Angelehnt an die Bezeichnerstrategie, jedes Einzelseitenobjekt mit einer URN zu versehen, kam die Leistungsvignette „URN granular“ in den Umlauf. Dieses Feature wurde von der Firma semantics GmbH[19] (Aachen) in Verbindung mit den Digitalisierungsprojekten der ULB Sachsen-Anhalt [00z] (z.B. Sammlung Ponickau) und von der Firma intranda als

---

[15] DigiZeitschriften: Bibliographische Info / als Beispiel: Anzeiger der Bibliothekwissenschaft. Jahrgang 1848/49 http://www.digizeitschriften.de/dms/met/?PPN=PPN340870265_0004-0005 – abgerufen am 14.04.2014

[16] Eine Monographie „Atlas selectus von allen Königreichen und Ländern der Welt“ unter der URL <http://digital.slub-dresden.de/werkansicht/dlf/1425/1/cache.off> mit der Band-URN: urn:nbn:de:bsz:14-db-id2859920663 – abgerufen am 14.04.2014

[17] SPO – Scripta Paedagogica Online am Beispiel aus der Zeitschrift Der katholische Jugendbildner - 6.1844, Seiten-URN: http://goobiweb.bbf.dipf.de/viewer/resolver?urn=urn:nbn:de:0111-bbf-spo-12120346 – abgerufen am 14.04.2014

[18] www.intranda.com

[19] http://www.semantics.de/

rückwärts kompatible METS-Datenerweiterungen vermarktet, so daß dem DFG-Viewer der profilierte Datenanteil weiterhin transparent dargeboten wurde.

Bei der Untersuchung der verschiedenen digitalen Sammlungen erkennt man, daß die Bereitstellung von URNs unterschiedlich gehandhabt wird. So zeigt die UB Kassel im „intranda viewer“ die PURLs an[20]. Es gibt jedoch über die OAI-PMH-Schnittstelle des Kasseler Viewers auch die Möglichkeit auf URNs zuzugreifen[21]. Bei einigen Beispielen endet die Vergabetiefe von URIs auf der Bandebene. Die Einzelseiten der Goobi-Projekte der SLUB Dresden [00aa] zeigen URL-Wechsel beim Blättern z.B. nach dieser Art:

- [URL]?id=5363&tx_dlf[id]=1425&tx_dlf[page]=16 dann folgend:
- [URL]?id=5363&tx_dlf[id]=1425&tx_dlf[page]=17 [siehe 00ab]

Man findet somit in der Sektion „Bildanzeige“ PURLs vor, deren Syntax sich an den Indizes der erfaßten Strukturobjekte orientiert (Digizeitschriften) oder die URN-NISS (siehe zum Begriff Seite 17) in die PURL einbeziehen (UB Kassel), wie das nachfolgende Beispiel zeigt:

- urn:nbn:de:hebis:34-02009112560193 = http://orka.bibliothek.uni-kassel.de/viewer/image/02009112560193/1/
- urn:nbn:de:hebis:34-02010031986915 = http://orka.bibliothek.uni-kassel.de/viewer/image/02010031986915/1/

Es ist einzuräumen, daß bei der Benutzung eines Webbrowsers die Bildanzeige und die Anzeige der Metadaten als naheliegendste und erste Kontaktstelle mit der URI-Variante als Kriterium für die URI-Typ-Einordnung hier willkürlich festgelegt ist. Eine tiefere Exploration, wie zum Beispiel über den METS-Download oder einer webbasierten OAI-PMH-Schnittstelle, ist ein folglich nachrangiges Kriterium für die soeben nur stich-probenartige Einordnung der Viewer in PURL- oder URN-Anzeiger.

---

[20] Beispiel: http://orka.bibliothek.uni-kassel.de/viewer/image/1366290520646/10/ – abgerufen am 14.04.2014

[21] OAI-PMH-Befehl zum Auflistung der Identifier einer bestimmten Kollektion zugehörig: http://orka.bibliothek.uni-kassel.de/viewer/oai?verb=ListIdentifiers&metadataPrefix=oai_dc&set=fotografie.historischefotosammlungdermurhardschenbibliothek

Kritik 4

Die hier nur angedeutete Variationsbreite und Verfahrensfreiheit bei dem Umgang mit URIs zeigt, daß das Gesamtbild uneinheitlich ist. Es fehlt an einem standardisierten Informationsschema der Web-Portale, was die Beschreibungstiefe und Lieferungsart mit Identifikatoren übersichtlich vermitteln könnte. Über die Anwendungsverteilung und Verbreitung der beiden Modelle Adress- versus Objektidentifizierung gibt es keine statistischen Daten.

4. URN granular bei Scripta Paedagogica Online (SPO) und ein Problem

Scripta Paedagogica Online (SPO) veröffentlicht über seinen METS-Viewer von der intranda GmbH das Ergebnis eines Migrationsprojektes. Er wurden ca. 243.050 Zeitschriftenartikel aus vorangegangenen Digitalisierungsprojekten, die in ca. 2.207 Zeitschriftenbänden von 166 Zeitschriften ab den Jahr 1766 erschienen sind, in das System Goobi.Production importiert. Die Artikeldatensätze der Migrationsquelle beinhalteten das übliche bibliographische Datenset. Hinzu kamen noch Metadaten, die über ein Skriptsystem eine ähnliche, jedoch eingeschränktere Viewer-Funktion ermöglichten. Eine Einschränkung bestand darin, daß die den Artikelobjekten zugehörigen Einzelseitenscans in Form eines Dateiintervalls gesammelt wurden. Im zugehörigen Datensatzfeld standen als Intervalldefinition eine Unter- und Obergrenze in Form von zwei Dateinamen, die mit Trennzeichen markiert waren, nach dem Schema: „/Verzeichnisstruktur/001.gif - 004.gif“

Jede dateisystemgemäße Einordnung von Elementen in dieses Intervall wurde in die Artikelanzeige aufgenommen. Ein 4-seitiges Dokument konnte demnach Dateien mit den Namen beinhalten[22]: 001, 002, 003, 004 oder 001, 001a, 001b, 002 (stets mit Suffix „.gif“). Damit konnten in das notierte Intervall „/Verzeichnisstruktur/001.gif – 004.gif“ ohne weiteres beispielsweise 20 Bilddateien mit entsprechenden Dateinamen gültig plaziert sein, wie z.B. 001, 001a, … 001g, 002, 003, 003a, 003e usw.

Im Migrationsprojekt von SPO, das den Umzug der digitalen Bibliothek vom alten System nach Goobi behandelte, wurde daraufhin das Verhältnis zwischen Materialobjekt Seite und Formalobjekt Seite stets kritisch bedacht.

[22] In dem früheren Viewer wurden die Dateinamen der Elemente des Seitenintervalls über eine Betragsfunktion umgewandelt. Das Modell sah vor, daß die erste Datei des Artikels einen Abstand vom Artikelbeginn "1" hat und so fort bis zur Obergrenze des Intervalls (Artikelende). Damit hatten die Einzelseiten jedoch aussageschwache URLs, weil diese Absolutwert-Bilddatei-Paarung keine tatsächliche Paginierung wiedergeben konnte.

4.1 Diskursobjekt Scanseite

Mit der Einzelseiten-URN in den von Goobi geschriebenen METS-Dateien sollte diese Identifizierungsgrenze des Dateiintervalls durchbrochen werden. Es stellt sich nun die Frage, wie diese Einermenge „Seite“, nämlich die eindeutig identifizierbare Abbildung eines Objekts, herrührend von einem Einzelarbeitsschritt an einem einzelbildgebenden Gerät, als Diskursobjekt behandelt werden soll. Ist diese Einermenge als eine Entität definiert, so werden Aussagen über diese Seite bzw. über die Informationen, die sich (nur) auf dieser Seite befinden mit der Objektrepräsentation z.B. in Form einer URN verknüpft und andernorts notiert. Die Motivation für diese Verknüpfungen wurden oben schon in den DFG-Praxisregeln zitiert. Es soll bei dem „zuverlässiges Arbeiten mit den bereitgestellten Quellen in wissenschaftlichen Kontexten“ [00g] die zuverlässige Identifizierung eine notwendige Bedingung sein. Neben den bereitgestellten Inhalten, die über diese Seitenabbildungen als zentrales Motiv der elektronischen Publikation gelten, ist eine weitere semantische Ebene Ort einer paradigmatischen Beziehung. Hierfür soll definiert sein: „Beziehungen zwischen Einheiten, die in ein und demselben Kontext auftreten können und sich in diesem Kontext gegenseitig ausschließen heißen PARADIGMATISCH.“ [...] „Paare von Einheiten, die in paradigmatischer Beziehung zueinander stehen, bilden eine OPPOSITION.“ [Wagn00a] Die hierfür genannten Einheiten sind „vorhandene Seite“ und „nicht vorhandene Seite“. Diese abstrakten Einheiten sollen dies verdeutlichen. Es kann nicht im selben Kontext eine Seite vorhanden und gleichzeitig dieselbe nicht vorhanden sein. Die Beziehung von körperlicher Überlieferungsform (Originalobjekt unter dem Scanner) und elektronischer Migrationsform (Digitalisat) im Hinblick auf Seitenexistenz ist nicht symmetrisch. Das Fehlen einer Seite in der körperlichen Überlieferungsform, hier genannt Original, führt zum Fehlen einer Seite im Digitalisat, jedoch nicht umgekehrt. Es besteht also nur im gerichteten Verhältnis $Original \rightarrow Digitalisat$ eine Wertverlaufsgleichheit insofern, als der durch Zuwachs manipulierte zahlenmäßige Seitenumfang des Originals den zahlenmäßigen Seitenumfangsemulation[23] des Digitalisats beeinflussen kann und nicht umgekehrt. Vorausgesetzt ist hierbei, daß alle Originalseiten 1:1 in das Digitalisat

A	B	$A \Rightarrow B$
f	f	w
f	w	w
w	f	f
w	w	w

Tafel 1

[23] Ein digitales Objekt hat keine Seiten, kann jedoch den seitenbezogenen Aufbau einer körperlichen Überlieferungsform logisch aufnehmen und in einem Sicht- und Navigationskontext emulieren.

übertragen werden. Man kann diese Logik auch in einer Implikation ausdrücken. Siehe dazu Tafel 1 auf Seite 38.

Es sei Digitalisat A und Original B . Die Implikation ist $A \Rightarrow B$ und drückt die hinreichende Bedingung aus: Schon wenn ein Digitalisat im elektronischen Bildstapel ist, hat dieses Digitalisat einen Originalvorfahren. Zugleich gilt hierfür die notwendige Bedingung: Nur wenn ein Digitalisat (A) einen Originalvorfahre (B) hat, darf dies im Stapel sein. Originalvorfahre (B) ist notwendige Bedingung, aber nicht ausreichend. Das Digitalisat (A) kann auch aus technischen oder rechtlichen Gründen dennoch im Bildstapel fehlen. Die Wahrheitswertetafel der Implikation $A \Rightarrow B$ ist hierfür wie folgt: Es ist nur der Fall falsch (f), wenn das Digitalisat A als wahr (w) deklariert ist, obwohl kein Original B vorhanden ist.

Es besteht jedoch in der zweiten Zeile der Wahrheitswertetafel 1 (S. 38) der Fall, daß kein Digitalisat vorhanden ist, jedoch ein Original vorliegt. Die Implikation leitet jedoch daraus „wahr“ ab. Für die gewünschte reale Kontrolle, was hier die Absicht ist, ist das nicht dienlich. Damit erkennt man noch nicht die fehlerbedeutende Situation, daß eine Digitalisierung fehlt. Eine Umkehrung der Implikation nach $B \Rightarrow A$ gäbe einen wahren Wert aus, wenn der Wert für das Original „falsch“ ist und für das Digitalisat „wahr“. Diese logische Schlußfolgerung ist für die reale Situation, wo eine bestimmte Kausalität herrscht, ebenfalls nicht dienlich. Die logische Beziehung der Konjunktion wäre naheliegend. Siehe Tafel 2.

A	B	$A \wedge B$
w	w	w
w	f	f
f	w	f
f	f	f

Tafel 2

Sie kann jedoch das definitive Fehlen einer Originalseite und das kausal damit zusammenhängende Fehlen eines Digitalisats nicht als „wahr“ ausdrücken.

Es soll deswegen ein weiteres abstraktes Objekt C eingeführt werden, so daß eine dreistellige Operation geschaffen wird. (Siehe Abb. 10) Der zusammengesetzte Ausdruck muß dann wahr sein, wenn der Wert von A, B und C jeweils „wahr“ ist und außerdem dann, wenn der Wert von A und von B jeweils „falsch“ ist, jedoch der Wert von C gleich „wahr“ ist. (nach [Varg70] S. 75) In allen anderen Fällen muß der zusammengesetzte

Ausdruck „falsch“ sein. **Es soll die Bedeutung des Objekt C definiert sein als „Digitalisierung(sversuch) tatsächlich geprüft und bestätigt“.**

4.1.1 Exkurs zum „Falsch“-Begriff

Neben dem natürlichsprachlichen Begriff „falsch“ wird in dieser Arbeit auch einer der Wahrheitswerte aus der zweiwertigen klassischen Logik mit „falsch“ bezeichnet. Der Wahrheitswert „falsch“ im logischen Sinne wird in den Wahrheitswertetafeln mit f gekennzeichnet.

Bei der Auseinandersetzung mit dem natürlichen Begriff „falsch“ geht es um die potentiell strittige Frage, ab wann eine digitale Abbildung eines körperlichen Informationsträgers falsch ist. Statt sich in die Fülle einer induktionistischen Tour d'Horizon zu begeben, sei es vorgezogen, einen der Digitalisierung übergeordneten Sachwaltungsbereich heranzuziehen, um dort bereits erarbeitete normative Aussagen zur Phänomenologie der Konvention des Falschbegriffs verwerten zu können.

Dieser übergeordnete Bereich ist die digitale Langzeitarchivierung (dLZA). Als normative Quelle für die dLZA ist die DIN-Norm 31644[24] bestimmend. Die Norm wird in den DFG-Praxisregeln [00g] in der Fußnote 46 neben dem Dokument „Kriterienkatalog vertrauenswürdige digitale Langzeitarchive“ [ScDA08] genannt. In letztgenanntem Band 8 der Reihe „nestor-Materialien“ können zwei Aussagen zur Profilierung einer DIN-gerechten Konvention dienlich sein: (a) *„Ein Informationsobjekt ist eine logisch abgegrenzte Informationseinheit. Ein Informationsobjekt kann (teilweise oder vollständig) durch digitale Objekte repräsentiert werden.“* (b) *„Authentizität bedeutet hier, dass das Objekt das darstellt, was es vorgibt darzustellen. Ein wichtiger Aspekt ist, dass das vorliegende Objekt von der angegebenen Quelle […] erstellt wurde.“* Es sei noch eine Aussage aus der DIN-Norm anzufügen, als diese Norm zur Kommentierung

[24] http://www.beuth.de/de/publikation/vertrauenswuerdige-digitale-langzeitarchivierung-nach-din-31644/169654635
Siehe auch nestor Kurzmeldung: Kommentarband zu DIN 31644 erschienen • http://www.langzeitarchivierung.de/Subsites/nestor/DE/Home/Kurzmeldungen/kommentarband.html

frei zur Einsicht stand: *„Maßstab für die Authentizität sind die signifikanten Eigenschaften des Informationsobjekts.“*

Eine gewinnende Auswertung dieses Sachwaltungsbereichs hat Grenzen. Am Eintrittpunkt in die dLZA erscheint das aufzunehmende Informationsobjekt stets als perfekt in seinem Staus quo. Lediglich in Verbindung mit der Authentizität wird die vermutete Bestimmung des Objektes in eine „signifikante Eigenschaft“ übersetzt. Dieser Rückverweis der spartenfremden Dokumentvalidierung in das Vorfeld der dLZA sollte zu einer Topologie der Vervollständigung anregen. Das Vervollständigen einer Abbildung der körperlichen Überlieferungsform sollte von einem endlichen Kern ausgehend jedes Regulativ durchbrechen, das revisions- und verbesserungsunfähige Endzustände einer dokumentarischen Bezugseinheit anstrebt. Die flexible Vervollständigung würde erlauben, daß Platzhalterabbildungen für fehlende körperliche Dokumentteile vorübergehend einspringen. Ebenso würden signifikante Lücken mit wieder aufgefundenem Quellenmaterial digital geschlossen werden und die Geschichte der Platzhalter außerdem bewahrt bleiben.

	A	B	C	$D = [(A \wedge B \wedge C) \vee (\sim A \wedge \sim B \wedge C)]$
1	w	w	w	w
2	w	w	f	f
3	w	f	w	f
4	w	f	f	f
5	f	w	w	f
6	f	w	f	f
7	f	f	w	w
8	f	f	f	f

Abb. 10: Wahrheitswertetafel für den zusammengesetzten Ausdruck D

Wird die Digitalisierung vollzogen, sind damit zwar Original und Digitalisat existent. Die Digitalisierung bekommt jedoch zunächst den C-Wert „falsch“. Erst wenn eine intellektuelle Prüfung, die durchaus mit dem Prozess der Digitalisierung ineinander fallen kann, erfolgt, entscheidet sich dadurch die Wertänderung entweder zu den Zeilen 1 oder 3 oder 5. Die Zeile 8 in Abb. 10 könnte man als kategorische „offene-Welt-Annahme“

betrachten: Zunächst existieren beliebig viele Lücken, denen nicht widersprochen wird bzw. die nicht falsifiziert sind.

Der Wert des zusammengesetzten Ausdrucks $(A \wedge B \wedge C) \vee (\sim A \wedge \sim B \wedge C)$ **(dies sei ab nun** D **)** kann durch irgendeine Systemschnittstelle veröffentlicht werden und jede „falsch"-Aussage durch D könnte kollaborativ geheilt werden, in dem der Fehler über ein abfragbares System lokalisierbar gemacht wird und - wenn möglich - ein nachgeholtes Digitalisat die Lücke füllt. Es sei D im Fall von „wahr" D und im Fall von „falsch" $\sim D$, d.h. „nicht D" bzw. $\neg D$.

Es besteht mit der dreistelligen Operation die Möglichkeit, z.B. in Zeile 7 den Verlust der Originalseite B als gegeben zu deklarieren. In Zeile 5 könnte das Digitalisat als falsch bzw. fehlerhaft deklariert worden sein. Der Zustand Zeile 1 wird damit zum Zustand Zeile 5 . Damit ist die intellektuelle Prüfung C verworfen. Ziel ist immer Zustand Zeile 1 oder Zeile 7. Alle anderen Zustände ergeben sich durch Änderungen bei A und B , ggf. auch bei C, wenn ein $C = w$ in Zeilen 1 und 7 aus irgendwelchen Gründen zu $C = f$ geändert wird (siehe Zeilen 2 und 8 von Abb. 10).

Mit dem zusammengesetzten Ausdruck D für „Digitalisierung vollständig" und „Digitalisierung unvollständig" kann man ein weiteres Modell konstruieren. Mit dem Modell sollen z.B. die Konzepte „vollständiger Heftartikel" und „unvollständiger Heftartikel" beschrieben werden. Das Fehlen einer Seite in einem Heftartikel drückt einen Fehler aus. Es muss dafür erwiesen sein, daß diese Seite dem Heftartikel zugehörig ist. Für die Konstruktion des Kontextes seien folgende Attribute und Werte einem Merkmalbündel zugeordnet und in nachfolgender Tabelle notiert.

Attribut	**Wert**	**Erklärung**
Elemente	{Digitalisate, Originale}	Lexikalische Kategorie, Kategorialsymbole A, B
Grundmenge	$H = \{h_1 ... h_n\}$	Intentionale Menge der Seitenelemente des dem Artikel übergeordneten Strukturelements mit der Vorschrift: Jede Seite soll nur einmal in der Menge sein.
Bildfunktion f^1	Interpretation des Artikelumfangs	Leitet Bildmenge von Grundmenge ab $f^1 = \{\langle h_1, j_1 \rangle, \langle h_2, j_2 \rangle, ... \langle h_n, j_n \rangle\}$
Bildmenge	$J = \{j_1 ... j_n\}$	Definierte Seiten unter f^1
Klasse	{Band, Heft, Artikel}	Lexikalische Kategorie, Kategorialsymbole E, F, G
Bildfunktion f^2	Werte von Ausdruck D {wahr, falsch} bzw. $D \vee \neg D$	Leitet Bildmenge von Bildmenge unter f^1 ab $f^2 = \{\langle j_1, d_1 \rangle, \langle j_2, d_2 \rangle, ... \langle j_n, d_n \rangle\}$

Tabelle 2: Merkmalbündel Strukturelement Zeitschriftenheftartikel

Wenn nun bei einer Überprüfung der Fall des Fehlens einer Seite in Kategorie A oder B und wahlweise dort in Kategorie E oder F oder G feststeht, kann in der Bildfunktion f^2 die vermutete Fehlseite mit j_n im n-ten 2-Tupel notiert sein. Es ist zunächst unerheblich, ob die Lücke sogar aus mehreren Seiten besteht. Ausschlaggebend ist die Belegung von d_n mit Zuordnung $\neg D$ im 2-Tupel. Sobald eine Aussage $\neg D$ gefunden wird, kann über eine zugreifbare Transitivität jedes Strukturelement, in dessen

Bildbereich sich dieses 2-Tupel befindet, ebenso als sich im kritischen Status befindend abgeleitet werden.

Mit dieser kontextbezogenen Ausformung des Diskursobjekts „Scanseite“ ist schon ein einfacher Baustein für eine „Application Ontology“, als „spezielle, auf eine konkret focussierte Domäne oder Aufgabe zugeschnittene Ontologie, die in der Regel eine Domain und/oder Task Ontologie spezialisieren“ [Sack11a , Folie 69] vorbereitet. „Der Begriff Ontologie ist [...] als äquivalent zum Begriff Wissensbasis zu verstehen und beschreibt schlicht [...] ein Dokument, welches Wissen einer Anwendungsdomäne modelliert.“ [Hitz08 , S.12] Für diese Anwendungsdomäne von Vollständigkeitsaussagen über publizierte Werke muß zu den stets wahren Aussagen einer XML-Sprache eine Ausdrucksmöglichkeit für die Ableitung eines „nicht-wahr“ eingeführt werden. Für die Unterscheidung WAHR - FALSCH benötigt man als Basis unbedingt wahre oder unbedingt falsche Aussagen. „An ontology is a (possibly named) set of axioms. Axioms are stated in an ontology language. If all axioms of an ontology are stated in the same ontology language, then the ontology as a whole is in that ontology language.“ [Vran10, S. 23] Mit der Ontologie werden auf die syntaktischen Objekte Bewertungs- bzw. Interpretationsbegriffe streng von ersteren logisch getrennt angewandt. Die Logikanwendung bezweckt hierbei das maschinell Machbare und nicht das grundsätzlich Denkbare. Es sollte also ein Teilgebiet des Wissens beschreibbar und maschinell entscheidbar gemacht werden wie:

- Im Quellmaterial fehlen Seiten
- Im Zielsystem sind fremde Seiten oder Dubletten
- Im Zielsystem fehlen Seiten

Man kann bei vielen Millionen digitalisierten Seiten in Ermangelung eines formellen Informationsaustauschs über Fehlerzustände der Original-Derivat-Beziehungen davon ausgehen, daß solche kritischen Digitalisate in den letzten Jahren Produktionskontrollen ungehindert passierten und in der Zukunft auch dies tun werden. Besonders anfällig sind dafür automatisierte Migrationsprojekte. Die terminologisch zu beschreibenden Beziehungen zwischen Papierseite und Scanseite in einer Ontologie werden jedoch nicht nur bestehende Fehler im Bestand beider Materialwelten, sondern auch nachgeordnete Wissensverluste anzeigen und vielleicht auch verhindern können. „According to the

Semantic Web ontology languages, ontologies do not include only terminological knowledge - definitions of the terms used to describe data, and the formal relations between these terms - but may also include the knowledge bases themselves, i.e. terms describing individuals and ground facts asserting the state of affairs between these individuals.“ [Vran10, S. 24]

4.2 Paradigmatische Objektrelationen

Digitalisierte Druckwerke kann man nach wie vor als Einheiten von einer analoger Quelle (nämlich das gedruckte Werk) und einem digitalem Derivat betrachten. In den digitalen Bibliotheken wird wenigstens implizit irgendwie verbürgt, daß es diese Einheit tatsächlich gibt oder zumindest gab, wenn das Original mittlerweile zerstört ist. Man könnte sehr wohl dieser Paarung in einem bestimmten Kontext eine Austauschbarkeit zusprechen.

$$\text{Diese Seite ist Bestandteil von einem} \left\{ \begin{array}{l} Band(Original) \\ Band(Digitalisat) \\ Heft(Original) \\ Heft(Digitalisat) \\ Artikel(Original) \\ Artikel(Digitalisat() \\ Abschnitt(Original) \\ Abschnitt(Digitalisat) \end{array} \right\}$$

Abb. 11: Kommutierende Elemente im Kontext formaler Seitenzugehörigkeit

Die Begriffe „sind hinsichtlich bestimmter invarianter Eigenschaften austauschbar, sie KOMMUTIEREN. […] Elemente, die in einem gegebenen Kontext austauschbar sind, stehen in PARADIGMATISCHER Beziehung zueinander; sie bilden ein PARADIGMA.“ [Wagn00a] Das kommutierende Verhältnis kann jedoch durch eine zusätzliche Bedingung eingeschränkt sein, in dem der Kontext verändert wird und das Paradigma dadurch auf seine Relativität hindeutet. Wenn man aussagt:

> „Diese Seite ist Bestandteil von einer bibliographischen Struktur und, allgemein genannt, Bestandteil von einem Strukturelement“,

so wird der Wahrheitsgehalt des Satzes vor „und“ nicht verändert. In diesem neuen Kontext bilden die Elemente aus Abb. 11 „ein (lexikalisches) Paradigma, dessen Elemente sich dadurch auszeichnen, daß sie jeweils dem Wort [Strukturelement] begrifflich untergeordnet sind.“ [Wagn00a] Der Kontext beschreibt, daß „Strukturelement“ die anderen Elemente logisch impliziert. Die Art und Weise der Unterordnung bezieht sich auf das Begriffliche, wobei gelten soll: „Ein BEGRIFF ist eine mentale Repräsentation, eine ‚Wissenseinheit‘, die Klassen von Objekten und Sachverhalten aufgrund ihrer invarianten Merkmale zu einem Ganzen zusammenfaßt.“ [Wagn00a] Als invariantes Merkmal kann man allen Begriffen in der Klammer von Abb. 11 den Behältercharakter zuordnen.

Wenn nun die Aussage gilt,

$$\text{Diese Seite ist Bestandteil von einer } \left\{ \begin{array}{l} Seite(Original) \\ Seite(Digitalisat) \end{array} \right\}$$

Abb. 12: Begriffsunterschiede Seite

So ist es sinnvoll zu fragen, ob im Kontext einer eindeutigen Referenzierbarkeit der Seite, die Seite selbst oder deren Behälter eine URI erhalten sollte. Die Papierseite besitzt hierbei schon eine gewisse natürliche Stabilität, was ihre Autentizität über das Selbstbehältnis betrifft. Das Digitalisat im System von Goobi liegt als Bilddatei hingegen in einer Schichtung von abstrakten Behältern, wobei einer dieser Behälter die URI stellvertretend für die Bilddatei (und damit für die digitalisierte Seite) hält.

Im Folgenden soll der Umstand untersucht werden, warum bereits in Goobi zugewiesene URNs für digitalisierte Einzelseiten nach einer Veränderung des zugehörenden Bildstapels (aufgrund von Korrekturen in Bezug auf die Anzahl der Bilddateien) in der METS-Datei zu anderen Objekten verweisen. Es handelt sich hier vermutlich um modellbezogene Grenzen in den Algorithmen der Anwendung, die von der Annahme herrühren, daß es nur perfekte Digitalisierungsvorgänge gibt. Dabei wird also angenommen, daß nur mit dem oben gebildeten abstrakten Ausdruck D = wahr (siehe Abb. 10) klassifizierte Digitalisate in die Präsentationsumgebung gebracht werden.

4.3 Korrekturszenario bei Scripta Paedagogica Online (SPO)

Es sei in einer beispielhaften Monographie mit einem Umfang von 34 Seiten (= 34 Bilddateien) die innere Struktur aus sechs Kapiteln gebildet. Das Kapitel wird als Strukturobjekt instanziiert (Dg1 D8). Es werden ihm Metadaten (Dg1 FG3), sowie drei Inhaltsdateien des Bildstapels zugeordnet (Dg1 JK4-6). Die Inhaltsdateien haben die Stapelindizes 24, 25 und 26 (Dg1 BH9) mit den Paginierlabels "108", "109", "110". Es handelt sich konkret um die Bilddateien im TIF-Format, wie in (Dg1 BG1) angegeben.

```
<mets:div TYPE="page" ID="PHYS_0024" DMDID="DMDPHYS_0024" ORDERLABEL="108" ORDER="24">
    <mets:fptr FILEID="FILE_0023"/>
</mets:div>
<mets:div TYPE="page" ID="PHYS_0025" DMDID="DMDPHYS_0025" ORDERLABEL="109" ORDER="25">
    <mets:fptr FILEID="FILE_0024"/>
</mets:div>
<mets:div TYPE="page" ID="PHYS_0026" DMDID="DMDPHYS_0026" ORDERLABEL="110" ORDER="26">
    <mets:fptr FILEID="FILE_0025"/>
</mets:div>
```

Abb. 13: Ausschnitt structMap(physical) Seiten von Kapitel 5 im internen XML-Format

In der validen METS-Datei des Viewers spiegelt sich dieser Sachverhalt so wieder:

```
<mets:div TYPE="page" ID="PHYS_0024" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205933" ORDERLABEL="108" ORDER="24">
    <mets:fptr FILEID="FILE_0023_PRESENTATION"/>
    <mets:fptr FILEID="FILE_0023_THUMBS"/>
    <mets:fptr FILEID="FILE_0023_DEFAULT"/>
    <mets:fptr FILEID="FILE_0023_MIN"/>
    <mets:fptr FILEID="FILE_0023_MAX"/>
</mets:div>
<mets:div TYPE="page" ID="PHYS_0025" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205945" ORDERLABEL="109" ORDER="25">
    <mets:fptr FILEID="FILE_0024_PRESENTATION"/>
    <mets:fptr FILEID="FILE_0024_THUMBS"/>
    <mets:fptr FILEID="FILE_0024_DEFAULT"/>
    <mets:fptr FILEID="FILE_0024_MIN"/>
    <mets:fptr FILEID="FILE_0024_MAX"/>
</mets:div>
<mets:div TYPE="page" ID="PHYS_0026" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205956" ORDERLABEL="110" ORDER="26">
    <mets:fptr FILEID="FILE_0025_PRESENTATION"/>
    <mets:fptr FILEID="FILE_0025_THUMBS"/>
    <mets:fptr FILEID="FILE_0025_DEFAULT"/>
    <mets:fptr FILEID="FILE_0025_MIN"/>
    <mets:fptr FILEID="FILE_0025_MAX"/>
</mets:div>
```

Abb. 14: Ausschnitt structMap(physical) Seiten von Kapitel 5 im validen METS

Das für das valide METS dienende Diagramm 3 unterscheidet sich zunächst vom Produktionsformat (Diagramm 1) lediglich dadurch, daß URNs (Dg1 AB3-8) in die structMap(Physical) übertragen werden (Dg3 BJ10-11).[25]

[25] Im Übrigen zeigt das Diagramm 3 nur eine Bilddateivariante (MIN = Minimalansicht) mit den zugehörigen JPEG-Bilddateien an, um das Diagramm nicht zu überfrachten.

Zu den Stapelindizes ist im METS-Profil des DFG-Viewers vorgegeben: „Das ORDER Attribut darf lediglich einen ganzzahligen Wert (Integer) enthalten, der auf Ebene der Seiten eindeutig sein muß. Für die Reihenfolge der Seiten sind einzig die Werte der ORDER Attribute ausschlaggebend; die Reihenfolge der Elemente bleibt unberücksichtigt. Für den Seiten untergeordnete Strukturelemente wird die Verwendung des ORDER Attributs ebenfalls empfohlen.“ [Funk09 , S. 15]

4.4 Korrekturfall

Man nehme an, es wird eine zuvor in der Quelle übersehene Seite zusätzlich gescannt und soll in den Bildstapel integriert werden. Damit wächst der Bildstapel insgesamt von 34 auf 35 Bilddateien an. Es gilt, daß die neue Scanseite mit "109b“ zu paginieren ist.

Zwei Korrekturmöglichkeiten bieten sich an:

Variante 1: Einfügen der Bilddatei an das Ende des Bildstapels; in dem Dateisystem als Datei "00000035.tif“ gespeichert.

Variante 2: Umbenennen der Bilddateien im Dateiverzeichnis ab Datei "00000026.tif“ bis "00000034.tif“ um ein Zähler höher, damit eine Lücke zwischen Bild 25 und Bild 27 geschaffen ist. Einfügen der neuen Bilddatei als "00000026.tif“ in den Bildstapel.

Zu Variante 1: Versucht man nun, diesem fünften Kapitel, das bisher drei Bilddateien umfaßte, das Bild zuzuordnen, so kann die systeminterne Dateisortierung nicht durchbrochen werden. Das höhergezählte Bild kann nicht zwischen Bild physisch 25 und 26 eingeschoben werden. Es kann lediglich als letztes Bild zum Artikel zugeordnet werden. Ebenso steht die 35. Bilddatei mit der Paginierung „109b“ im linken Listenfenster der Abb. 15 ganz unten eingeordnet zur Verfügung.

Abb. 15: Zuweisen der physischen Seitenobjekte zum logischen Strukturelement Kapitel

Sowohl in der Bandansicht als auch in der Kapitelansicht ist diese „zugehörige Seite“ stets hinten angestellt über die Seitennavigation des Viewers erreichbar.

Bei der *Variante 2* schafft der Workaround einer Dateiumbenennung die gewünschte Platzierung der Bilddatei in die Dateigruppe des Kapitels. (Dg2 JK3-7)

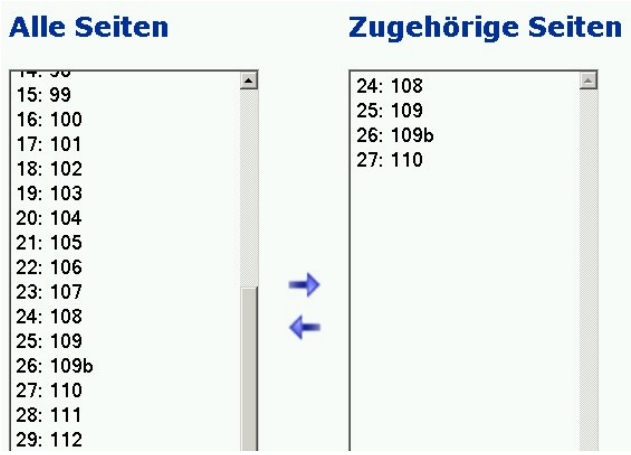


Abb. 16: Paginierungszuweisungen für Strukturelemente

Vergleicht man die METS-Dateien der Varianten, so ist festzustellen, daß die URN stets an dem Ergebnis einer mathematischen Betragsfunktion gebunden wird. Der Betrag ist der Abstand der Bilddatei im Stapel zum Beginn des Stapels, den man als 0 bezeichnen

kann. Die Datei 00000001.tif hat demnach den Abstand 1 zum Beginn des Bildstapels, die Datei 00000026.tif hat den Betrag 26. Diese Betragsfunktion wird auf den Identifier der beinhalteten -Elemente der METS-Sektion <mets:structMap TYPE=„PHYSICAL“> angewendet. Dies geschieht als globale Funktion stets, wenn der Bildstapel um Bilder ergänzt oder verringert werden muß. Die zwingende Prozedur ist in Goobi.Production mit dem Befehl „Paginierung anhand der Images einlesen“ vorzunehmen. Die nachfolgenden Abbildungen demonstrieren den Vergleich für die Situationen vor und nach der Korrektur.

Zunächst wird in Abb. 17 der 34-Bilder-Stapel gezeigt. Die mit "110" paginierte Seite in dem mit „PHYS_0026" identifizierbaren Element die URN mit den letzten Ziffern "56" und ist die letzte Seite des Kapitel 5. (Dg1 A5) zeigt die Seiten-URN-Zuweisung. (Dg1 E10) – zusammen mit (Dg1 E2) – zeigt die Verbindung des Elements aus der physischen structMap mit der Datei. Im Vergleich zu Abb. 13 sind die Repräsentanten der Seitenobjekte schon mit URNs bestückt, die in diesem Szenario als veröffentlicht zu betrachten sind.

```
- <mets:div TYPE="page" ID="PHYS_0024" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205933" ORDERLABEL="108" ORDER="24">
      <mets:fptr FILEID="FILE_0023_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0023_THUMBS"/>
      <mets:fptr FILEID="FILE_0023_DEFAULT"/>
      <mets:fptr FILEID="FILE_0023_MIN"/>
      <mets:fptr FILEID="FILE_0023_MAX"/>
  </mets:div>
- <mets:div TYPE="page" ID="PHYS_0025" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205945" ORDERLABEL="109" ORDER="25">
      <mets:fptr FILEID="FILE_0024_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0024_THUMBS"/>
      <mets:fptr FILEID="FILE_0024_DEFAULT"/>
      <mets:fptr FILEID="FILE_0024_MIN"/>
      <mets:fptr FILEID="FILE_0024_MAX"/>
  </mets:div>
- <mets:div TYPE="page" ID="PHYS_0026" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205956" ORDERLABEL="110" ORDER="26">
      <mets:fptr FILEID="FILE_0025_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0025_THUMBS"/>
      <mets:fptr FILEID="FILE_0025_DEFAULT"/>
      <mets:fptr FILEID="FILE_0025_MIN"/>
      <mets:fptr FILEID="FILE_0025_MAX"/>
  </mets:div>
- <mets:div TYPE="page" ID="PHYS_0027" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205962" ORDERLABEL="111" ORDER="27">
      <mets:fptr FILEID="FILE_0026_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0026_THUMBS"/>
      <mets:fptr FILEID="FILE_0026_DEFAULT"/>
      <mets:fptr FILEID="FILE_0026_MIN"/>
      <mets:fptr FILEID="FILE_0026_MAX"/>
  </mets:div>
```

Abb. 17: Ausschnitt structMap(physical) Seiten von Kapitel 5 im internen XML-Format mit URNs

Die XML-Sequenz zu (Dg1 E2) ist:

```
<mets:file ID="FILE_0025" MIMETYPE="image/tiff">
    <mets:FLocat xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="file:///opt/digiverso/goobi/metadata/3610/images/test_319141853_tif/00000026.tif"
        LOCTYPE="URL"/>
</mets:file>
```

Im 35-Bilder-Stapel zeigt sich der Versatz der Seiten-URN bzw. Seiten-URI.

```
- <mets:div TYPE="page" ID="PHYS_0024" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205933" ORDERLABEL="108" ORDER="24">
      <mets:fptr FILEID="FILE_0023_MIN"/>
      <mets:fptr FILEID="FILE_0023_MAX"/>
      <mets:fptr FILEID="FILE_0023_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0023_THUMBS"/>
      <mets:fptr FILEID="FILE_0023_DEFAULT"/>
  </mets:div>
- <mets:div TYPE="page" ID="PHYS_0025" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205945" ORDERLABEL="109" ORDER="25">
      <mets:fptr FILEID="FILE_0024_MIN"/>
      <mets:fptr FILEID="FILE_0024_MAX"/>
      <mets:fptr FILEID="FILE_0024_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0024_THUMBS"/>
      <mets:fptr FILEID="FILE_0024_DEFAULT"/>
  </mets:div>
- <mets:div TYPE="page" ID="PHYS_0026" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205956" ORDERLABEL="109b" ORDER="26">
      <mets:fptr FILEID="FILE_0025_MIN"/>
      <mets:fptr FILEID="FILE_0025_MAX"/>
      <mets:fptr FILEID="FILE_0025_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0025_THUMBS"/>
      <mets:fptr FILEID="FILE_0025_DEFAULT"/>
  </mets:div>
- <mets:div TYPE="page" ID="PHYS_0027" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205962" ORDERLABEL="110" ORDER="27">
      <mets:fptr FILEID="FILE_0026_MIN"/>
      <mets:fptr FILEID="FILE_0026_MAX"/>
      <mets:fptr FILEID="FILE_0026_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0026_THUMBS"/>
      <mets:fptr FILEID="FILE_0026_DEFAULT"/>
  </mets:div>
- <mets:div TYPE="page" ID="PHYS_0028" CONTENTIDS="urn:nbn:de:0111-bbf-spo-14205977" ORDERLABEL="111" ORDER="28">
      <mets:fptr FILEID="FILE_0027_MIN"/>
      <mets:fptr FILEID="FILE_0027_MAX"/>
      <mets:fptr FILEID="FILE_0027_PRESENTATION"/>
      <mets:fptr FILEID="FILE_0027_THUMBS"/>
      <mets:fptr FILEID="FILE_0027_DEFAULT"/>
  </mets:div>
```

Abb. 18: Ausschnitt structMap(physical) Seiten von Kapitel 5 im internen XML-Format mit URNs nach Einfügen neuer Seite 109b

Die im 34-Seiten-Dokument mit "110" paginierte Schlußseite des Kapitel 5 hat nicht wie in Abb. 17 die URN „…56“ behalten, sondern die URN „…62“ bekommen.

```
<mets:smLink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:from="LOG_0004" xlink:to="PHYS_0024"/>
<mets:smLink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:from="LOG_0005" xlink:to="PHYS_0024"/>
<mets:smLink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:from="LOG_0005" xlink:to="PHYS_0025"/>
<mets:smLink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:from="LOG_0005" xlink:to="PHYS_0026"/>
<mets:smLink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:from="LOG_0005" xlink:to="PHYS_0027"/>
<mets:smLink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:from="LOG_0006" xlink:to="PHYS_0028"/>
<mets:smLink xmlns:xlink="http://www.w3.org/1999/xlink" xlink:from="LOG_0006" xlink:to="PHYS_0029"/>
```

Abb. 19: StructLink-Abschnitt letzte Seite des Kapitel 5

Beim Vergleich von Diagramm 1 und 2 in (Dg1 AB3-8) und (Dg2 AB3-8) ist die Zuordnung der URN an die Seitenobjekte der physischen Struktur per Betragsfunktion[26] ersichtlich. Die URN verbleibt also an der 26. Stelle im Bildstapel.

[26] Siehe zur Betragsfunktion Seite 49

```
<mets:dmdSec ID="DMDLOG_0000">
<mets:dmdSec ID="DMDLOG_0001">
<mets:dmdSec ID="DMDLOG_0002">
<mets:dmdSec ID="DMDLOG_0003">
<mets:dmdSec ID="DMDLOG_0004">
<mets:dmdSec ID="DMDLOG_0005">
<mets:dmdSec ID="DMDLOG_0006">
<mets:dmdSec ID="DMDPHYS_0000">
<mets:dmdSec ID="DMDPHYS_0001">
<mets:dmdSec ID="DMDPHYS_0002">
<mets:dmdSec ID="DMDPHYS_0003">
<mets:dmdSec ID="DMDPHYS_0004">
<mets:dmdSec ID="DMDPHYS_0005">
<mets:dmdSec ID="DMDPHYS_0006">
<mets:dmdSec ID="DMDPHYS_0007">
<mets:dmdSec ID="DMDPHYS_0008">
<mets:dmdSec ID="DMDPHYS_0009">
<mets:dmdSec ID="DMDPHYS_0010">
<mets:dmdSec ID="DMDPHYS_0011">
<mets:dmdSec ID="DMDPHYS_0012">
<mets:dmdSec ID="DMDPHYS_0013">
<mets:dmdSec ID="DMDPHYS_0014">
<mets:dmdSec ID="DMDPHYS_0015">
<mets:dmdSec ID="DMDPHYS_0016">
<mets:dmdSec ID="DMDPHYS_0017">
<mets:dmdSec ID="DMDPHYS_0018">
<mets:dmdSec ID="DMDPHYS_0019">
<mets:dmdSec ID="DMDPHYS_0020">
<mets:dmdSec ID="DMDPHYS_0021">
<mets:dmdSec ID="DMDPHYS_0022">
<mets:dmdSec ID="DMDPHYS_0023">
<mets:dmdSec ID="DMDPHYS_0024">
<mets:dmdSec ID="DMDPHYS_0025">
<mets:dmdSec ID="DMDPHYS_0026">
<mets:dmdSec ID="DMDPHYS_0027">
<mets:dmdSec ID="DMDPHYS_0028">
<mets:dmdSec ID="DMDPHYS_0029">
<mets:dmdSec ID="DMDPHYS_0030">
<mets:dmdSec ID="DMDPHYS_0031">
<mets:dmdSec ID="DMDPHYS_0032">
<mets:dmdSec ID="DMDPHYS_0033">
<mets:dmdSec ID="DMDPHYS_0034">
<mets:dmdSec ID="DMDPHYS_0035">
<mets:fileSec>
```

Abb. 20: Erweiterter Bildstapel (geschlossene XML-Elemente) auf 35 Seiten im Produktionssystem. Zu DMDPHYS_0000 siehe [27]

Die URN „...56“ verbleibt am Objekt mit der ID „DMD_PHYS_0026“ (Dg2 AB5-6). Im -Element der structMap(physical) (Dg2 FG7) befindet sich ein -Element mit der Attribut-ID „PHYS_0027“ (Dg2 FG10). Es wird per <fptr>-Element die Attribut-FILEID mit dem Wert „FILE_0026“ (Dg2 FG11) angebunden und damit der Verweis auf die Bilddatei vorbereitet. In einem XML-Element <fileGrp> (Dg2 E4) ist das Verweisungsziel von PHYS_0027, nämlich die ID „FILE_0026“ notiert. Von dort wird auf die Datei „00000026.tif“ verwiesen.

In den Diagrammen werden die oval umrandeten Paginierungen an den URN-Symbolen und Dateinamen (Dg1, Dg2 und Dg3 jeweils Zeile 1) lediglich dazu genutzt, auf die wahren Bildinhalte der Seiten hinzuweisen. In (Dg3 EJ10-11), eine Abbildung der validen METS-Datei aus dem Viewer, wird das gefärbte Paar im Zustand des Getrenntseins nochmal dargestellt. Die Seite und die eigentlich dazugehörende URN haben die gleiche Farbe.

4.5 Konsequenzen des URN-Versatzes

Sobald die veröffentlichte URN von Nutzern als Referenz in weiteren Systemen genutzt wird, ist die Manipulation des Bildstapels bei dem jetzigen technischen Stand des Goobi-Systems offensichtlich problematisch. Das Projekt SPO wird deshalb die 2.207 digitalisierten und bereits veröffentlichten Bände manuell auf eventuelle Fehler überprüfen und ggf. Stellvertreterdateien für den Fall von Lücken im Original in den

[27] „Innerhalb des physischen <structMap> Elements wird die Seitenstruktur durch Elemente wiedergegeben, die einem obersten Element untergeordnet sind. Dieses oberste Element umfasst die Seiten, die die bibliographische Einheit repräsentieren. Daher muss dessen TYPE Attribut immer den Wert ‚physSequence‘ besitzen.“ [Funk09] Bei dem seitenbasierten Dokumentenmodell von SPO sind das in der Regel die Jahrgangsbände von Zeitschriften oder Buchbände.

Bildstapel stellen, wenn nicht sofort nachgescannt werden kann. Diese „Bild-Dummies“ weisen mit einem Pixelmuster in Textform darauf hin, daß diese Seite noch nicht digitalisiert werden konnte, weil das Original (noch) fehlt. Erst wenn diese Prüfung am Band abgeschlossen ist, werden die Seiten-URNs des Bandes im Präsentationssystem nachgereicht. Mit den Platzhalterdateien für fehlende Quellseiten kann man einen kontrollierten Status D nach dem oben beschriebenen Modell aus Kapitel 4.1 vorbereiten. Wenn die Quellseite als verloren gilt, kann der Status wie in Zeile 7 von Abb. 10, das ist D = wahr, vergeben werden.

Die Bildstapelkontrolle ist ein aufwendiger manueller Validationsprozeß, der die URN-Veröffentlichung hinauszögert. Falls eine Goobi nutzende Institution eine Software-erweiterung in Auftrag gibt, damit bei Änderungen am Bildstapel die URI-Bild-Paarungen beibehalten werden, ist die historisch gewachsene Streuung von gebrochenen Verlinkungen damit nicht automatisch rückgängig zu machen.

Die PURL-Version zur Kodierung eines Seitenidentifiers ist mit dem gleichen Problem konfrontiert. Die in Kapitel 3.3 besprochene Verteilung von URN- oder PURL-Adressierung auf Seitenebene baut auf derselben Technik der Bildstapelindex-Referenz auf (vgl. Abb. 17 und Abb. 18). Es sind also nicht nur die früh an einer Seiten-URN interessierten Bibliotheken betroffen, sondern im Prinzip alle. Es ist in Ermangelung statistischer Daten nicht zu ermitteln, wie viele Goobi-Bibliotheken ältere Systeme, wie bei „Scripta Paedagogica Online“, inklusive einer gewissen „Bildstapelblindheit“[28] nach Goobi migriert haben und sich mit ebensolchen Problemen konfrontiert sehen.

Die Vermutung, daß diese Lücken doch selten vorkommen, können die Notwendigkeit einer Erfassung oder Korrektur der Lücken nicht entkräften. Über diese Bild-URI-Paare ist als Leistungsmerkmal der Forschungsdatengenerierung eine zusätzlich verlinkte Datenschicht, nämlich die Ebene der Volltextinhalte (über OCR gewonnen), zu schützen.

[28] Es sei hier das Dateiintervall des alten Systems von SPO auf Seite 3737 angesprochen.

5. Erfüllende Aussagen

Das bisher beschriebene XML-basierte semi-formale Sprachkonstrukt und Sprachschema bedient sich eines strukturierten Vokabulars. Es läßt sich damit irgend etwas Beliebiges definieren. „Unter einer DEFINITION versteht man die genaue Abgrenzung eines Begriffes innerhalb eines größeren Zusammenhanges unter Verwendung anderer Begriffe (EXPLIZITE Definition).“ [Wagn00b] Diese geteilten Definitionen werden in der Anwendungsdomäne von Goobi sowohl für den Programmcode, als auch für das METS-Profil des DFG-Viewers und für die Regelsatzgestaltung zwischen Menschen ausgehandelt und angewandt. Für die Algorithmen des elektronischen Systems sind die semantischen Inhalte jedoch nicht zugänglich und somit nicht verarbeitbar. „The new standard of XML is a big improvement but can still support communications only in cases where there is a priori agreement on the vocabulary to be used and on its meaning.“ [AnHa08] In diesen „Agreements“ werden Ideen für Sachverhalte zum Zwecke der Wiedererkennung formuliert. Es gibt die Idee des „Kapitels“ und diese Idee ist wiedererkennbar, auch wenn die konkreten Kapitelobjekte unterschiedlich groß sind. Ebenso gibt es z.B. die Idee des „FLocat“ mit der Definition: „METS provides the ability for content either to be stored within the METS file itself or stored externally in another file and referenced. For this example, we will store all of the content externally and reference it using the <FLocat> sub-element of the <file> element“ [Digi10] Für Sender und Empfänger der Information zu <FLocat> ist die semi-formale Beschreibung des Sachverhaltes verarbeitbar, sofern sie in folgender Syntax vorliegt:

```
<mets:file ID="FILE_0025_MIN" MIMETYPE="image/jpeg">
      <mets:FLocat xmlns:xlink="http://www.w3.org/1999/xlink"
      xlink:href="http://goobiweb.bbf.dipf.de/viewer/content/319141853/6
00/0/00000026.jpg"
      LOCTYPE="URL"/>
</mets:file>
```

5.1 Kode-Kohorten in Tunnelgängen

Es gibt in diesem von Goobi genutzen universellen Datenmodell, [EnFu00b] , in der Art eines seitenbasierten Transportformats im METS/MODS-Standard, "gemeinsame Symbole und Begriffe (Syntax), [eine] Übereinkunft bzgl. deren Bedeutung (Semantik), [eine] Klassifikation von Begriffen (Taxonomie) [und auch] Assoziationen und

Vernetzungen von Begriffen (Thesauri) [sowie] Regeln und Wissen darüber, welche Vernetzungen zulässig und sinnvoll sind (Ontologien)". [Sack11a , Folie 10] Man verbindet ausdrücklich die Semantic Web Technlogien mit dem Erschaffen von Ontologien. "(A)n ontology, is a catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D." [Sowa00] Das reicht jedoch nicht aus. John F. Sowa setzt fort: "An uninterpreted logic, such as predicate calculus, conceptual graphs [...] is ontologically neutral. It imposes no constraints on the subject matter or the way the subject may be characterized. By itself, logic says nothing about anything, but the combination of logic with an ontology provides a language that can express relationships about the entities in the domain of interest." [Sowa00]

Erst wenn die Logik mit terminologischem Wissen kombiniert wird und damit die Ontologieanwendung explizit formuliert werden kann, sind auch erfüllende Aussagen möglich. Dabei ist nicht der natürlichsprachige Begriff „erfüllend“ das Motiv, sondern es geht um Aussagen aus der Logik heraus, die formelbezogen als „erfüllend“ entschieden, explizit berechnet werden können. Das Explizite nimmt Thomas R. Gruber in die Definition auf: „An ontology is an explicit specification of a conceptualization.” [Grub93] Diese Definition wird oft (und in seiner Authentizität m.E. nicht nachweisbar) zitiert mit dem Partizip Perfekt vor Konzeptualisierung „of a *shared* conceptualization“. Das Semantic Web Paradigma nimmt diese „shared“-Dimension in die globale Ausdehnung des World Wide Web auf und möchte die Inseln der Kode-Kohorten und Datensilos vernetzen.

Mit Entlehnung von Noam Chomsky soll folgendes abstraktes Modell als Indikator für eine geschlossene Anwendungsdomäne stehen: Es besteht „ein idealer Sprecher-Hörer, der in einer völlig homogenen Sprachgemeinschaft lebt, seine Sprache ausgezeichnet kennt und bei der Anwendung seiner Sprachkenntnis in der aktuellen Rede von solchen grammatisch irrelevanten Bedingungen wie
— begrenztes Gedächtnis
— Zerstreutheit und Verwirrung
— Verschiebung in der Aufmerksamkeit und im Interesse
— Fehler (zufällige oder typische)
nicht affiziert wird.“ [Chom70 , S. 14]

Die damit angenommene Sprachkompetenz unterscheidet Chomsky - ebenso wie Ferdinand de Saussure - von der Sprachverwendung (performance). „Die Unterscheidung, die ich hier vermerke, ist verwandt der Saussureschen Trennung in *langue — parole*; es ist jedoch notwendig, von Saussures Begriff der *langue* als lediglich einem systematischen Inventar von Einheiten abzugehen und zurückzugehen auf das Humboldtsche Verständnis der zugrunde liegenden Kompetenz als einem System generativer (‚erzeugender') Prozesse. (Diskutiert in Chomsky (1964)).“ [Chom70 , S. 14] In der Ausgabe von 1975 (statt 1964) schreibt Chomsky: „[...] we have the Humboldtian view that 'man muss die Sprache nicht sowohl wie ein todtes Erzeugtes, sondern weit mehr wie eine Erzeugung ansehen' (1836, § 8, p. LV). [...] The notion of 'form' as 'generative process' underlies Humboldt's entire account of the nature of language and of the use and acquisition of language, and constitutes perhaps his most original and fruitful contribution to linguistic theory.” [Chom75 , S. 17]

Das XML-Datenmodell bietet jedoch nur – metaphorisch gesprochen – Punkt-zu-Punkt-Verbindungen (Tunnel) von maschinellen idealen Sprechern-Hörern an. Für diese Sprecher-Hörer (Sender-Empfänger) muß eine explizit konstruierte Kodeüberlappung für das mapping der Schemata und Vokabularien konstruiert werden.

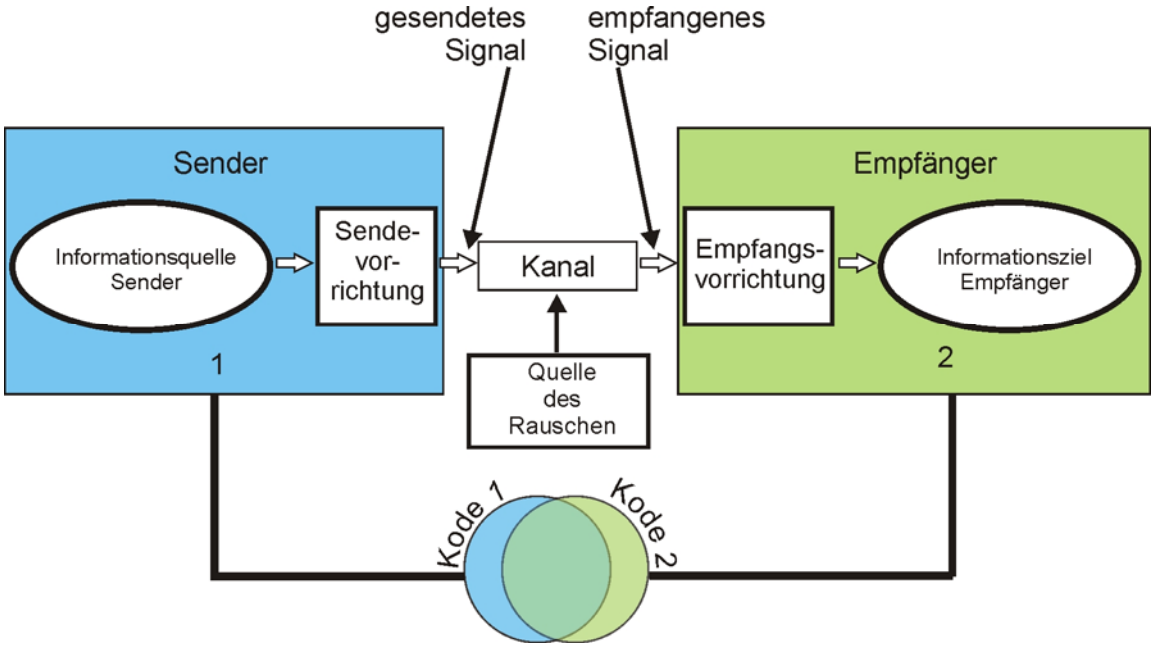


Abb. 21: Kommunikationsmodell mit sich überlappenden Kodes

Sobald unterschiedliche Begriffe und Konzepte mit eigenen, von einander unterschiedlichen Schemata vorliegen, sind für den Informationsaustausch Übersetzer nötig. Entsprechend müssen bei dem gegenseitigen Datenaustausch von n verschiedenen System n^2 Übersetzer hergestellt werden. Im semiotischen Verständnis sind XML-Dokumente k omplexe Zeichen. Die Bedeutung dieser Zeichen sind lediglich implizit, weil mit allen XML-Bestandteilen nur die Beziehung zwischen Zeichen zu anderen Zeichen, also lediglich die Dimension der Syntaktik durchdrungen wird. „XML is nothing more than a syntax. You need a metalanguage vocabulary to be able to use XML to record business domain information in such a way that any business can be documented, and RDF provides this capability." [Powe03] Eine Beschleunigung der Erzeugung von Sprache auf semantischer "lebendiger" Ebene - weitab von einem "todten Erzeugnis" (Humboldt) - ergibt sich schon dann, wenn diese syntaktischen Verwicklungen und Entwirrungen nicht mehr das Hauptgeschäft der Forschungsumgebungen sind.

5.2 Im Team: Parser und Reasoner

Es ist also zeitaufwendig, schwierig und risikoreich, was die mapping-Fehler oder Äquivalenzbarrieren betrifft, auf der Basis von komplizierten Hierachieverschachtelungen in XML-Dokumenten Informationen zwischen Systemen auszutauschen und demzufolge kann auch dem einen, eigenen System ein weiterer „fremder" Kommunikationspartner gegenübergestellt werden: die noch nicht realisierte, veränderte Systemgeneration als Nachfolgeschaft des eigenen Systems, über die man neue Sachverhalte speichern möchte, die auch noch neue Überlappungen über bestehende hierarchische Strukturen mitbringen können. Anlaß zu Überlappungen von Stukturen gäbe es z.B. bei der Absicht, Seitenkollektionen zusammenzustellen, die eine Reihe von Abbildungen eines Grafikers über Zeitschriftenbände hinweg präsentieren.

Ein weiterer Sachverhalt kann im gegenwärtigen Sytsem ebenfalls nicht abgebildet werden, der eine Fortsetzung eines Artikels im nachfolgenden Zeitschriftenheft darstellt. Nachfolgend in Abb. 22 bilden die schwarzen Ellipsen und weißen Quadrate die Artikelseiten zweier Artikel auf drei Strukturebenen ab. Auf Bandebene ist der gesamte Bildstapel abgebildet.

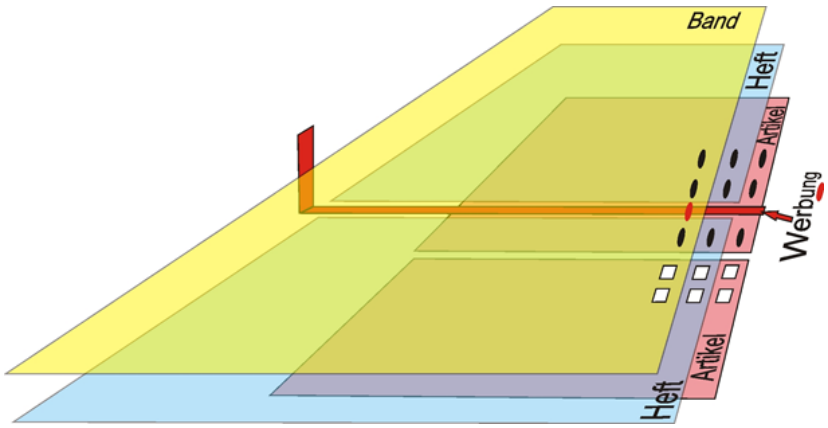


Abb. 22: Strukturelle Überlappungen bei Artikelfortsetzung im Folgeheft

Auf Heftebene ist die Verteilung des einen Artikels (Ellipsen) auf zwei Hefte vermerkt, wobei noch eine Werbeseite (rote Ellipse) zwischen den Heften eingebunden dargestellt ist. Die Artikelebene zeigt die implizite Logik der pragmatischen Zusammengehörigkeit der Artikelseiten. Im METS-Viewer sind diese Seiten (schwarze Ellipsen) jedoch nicht zusammengehörig in einer möglicherweise gewünschten "Gesamtartikelsicht" zu präsentieren. Zudem sollte diese Gesamtartikelansicht ggf. auch die Werbeseite ausblenden. Nach dem Goobi-Datenmodell kann ein Strukturelement, wie zum Beispiel ein Artikel, nicht zwei Elternelemente haben. Es müßten also aus diesem auf zwei Hefte verteilten Artikel ebenso, wie in der bibliographischen Formalerschließung, zwei Datensätze angelegt werden, die aufeinander verweisen. Das ist durch explizites Markup und anderweitiger Strukturierung nicht möglich. Die linguistische Sicht auf den gesamten Artikeltext wird auf zwei manuelle Ansätze gezwungen.

Die Überwindung der strengen XML-Hierarchieregel ist im METS-Dokument als Metamarkup angelegt, und daran ist die Prozeßlogik der Goobi Produktions- und Präsentationssoftware gebunden. Das Metamarkup nutzt die Technik der Fragmentierung von sich überlappenden XML-Sequenzen in mehrere – darin wieder hierarchisch strukturierte – Unterzonen. „The simplest type of non-hierarchical structure is the union of multiple structures, each of which is hierarchical.“ [Dero04] Über ID/IDREF-Mappings wird das verstreute Metamarkup zu einem logischen Ganzen wieder zusammengeführt, wie es z.B. in den TEI P5 Guidelines im Kapitel "20.3 Fragmentation and Reconstitution of Virtual Elements" diskutiert wird. [00ac]

Als ein unbemerkt weit verbreitetes Beispiel für die technische Bewältigung von Strukturüberlappungen mit XML-Syntax nennen [IoPV10] die sehr potenten Werkzeuge der Textverarbeitungsprogramme von Open Office (ODT-Format) und von Microsoft (DOCX-Format) und auch OOXML[29] (Microsoft Office 2007). [IoPV11][30] Mit diesen Programmen sind alle erdenklichen Änderungen am Textdokument, von mehreren Autoren in einem change-tracking-Modus aufgezeichnet, so daß Hinzufügungen und Löschungen zu jeder Zeit nachvollziehbar und ggf. reversibel sind.

Im Vergleich zu solch hochflexiblen Dokumenten sind die METS-Dokumente, die der DFG-Viewer validiert, relativ einfach strukturiert, deren wenige syntaktische Landmarken auch einfach abgefragt werden könnten und an denen in der Publikationsphase wenige Veränderungen vorgenommen werden. Iorio, Peroni und Vitali stellen in [IoPV10] über das Projekt EARMARK (Extremely Annotational RDF Markup) [Pero00a] zwei Abfragecodes (XPath und SPARQL) an dasselbe Diskursobjekt gerichtet gegenüber. "Our goal here is to show that complex overlapping data structures can be disentangled into EARMARK assertions that can be processed and queried in a very simple way." [IoPV10]

```
for $id in // @text:id [../ text:insertion //(dc:creator
[. = 'Angelo Di Iorio'] | @office:chg-author
[. = 'Angelo Di Iorio'])] return // text:p// text ()
[( preceding-sibling::text:change-start [1]
[ @text:change-id = $id] and following-sibling::text:change -end [1]
[ @text:change-id = $id ]) or
ancestor::text:changed-region/@text:id = $id]
```

Abb. 23: XPath-Abfrage aus: Iorio, Peroni und Vitali (2010) [IoPV10]

[29] Microsoft implementiert OOXML „strict“ in Office 2013 | heise online. URL http://www.heise.de/newsticker/meldung/Microsoft-implementiert-OOXML-strict-in-Office-2013-1668574.html. - abgerufen am 14.04.2014

[30] „ODT uses milestones and standoff markup for insertions and deletions, respectively, and also relies on standoff markup for annotations about the authorship and date of the change.” […] “The OOXML format (JTC1/SC34WG4, 2008) (the XML-based format used by Microsoft Office 2007 and standardized by ISO in 2008), on the other hand, uses a form of segmentation to store change-tracking information across all previous elements involved.” [IoPV11] S. 1699

```
SELECT ?r WHERE {
?r a text:insertion ; dc:creator "Angelo Di Iorio" }
```

Abb. 24: SPARQL-Abfrage aus: Iorio, Peroni und Vitali (2010) [IoPV10]

Die einfache Struktur und Stablität von Goobi-METS/MODS-Dokumenten könnte auch ein ideales Objekt für ein Kooperationsmodell sein, in dem Techniken des Semantic Web über den METS-Transportlayer von Goobi gelegt werden. Die Transportrolle des „page turners“ wird durch Goobi gut ausfüllt und ist weithin akzeptiert. Bevor im Kapitel 6 die Kooperation diskutiert wird, soll im folgenden Kapitel ein kleines Teil dieser neuen semantischen Schicht zur Modellierung eines Diskursobjekts „Seite“, das in Kapitel 4.1 aussagenlogisch beschrieben wurde, vorgestellt werden.

5.2.1 Prädikatenlogische Resolution

Zurückgreifend auf den abstrakten Begriff D im Kapitel 4.1 auf Seite 42 soll ein Bereich der Zustände in der aussagenlogischen Wahrheitswertetafel von Abb. 10 als Beispiel herausgenommen werden und in der Prädikatenlogik, dargestellt werden. „In der Aussagenlogik war es z.B. nicht möglich, auszudrücken, das gewisse ‚Objekte‘ in gewissen Beziehungen stehen; daß eine Eigenschaft *für alle* Objekte gilt, oder daß ein Objekt mit einer gewissen Eigenschaft *existiert.*“ [Schö00] Mit Hilfe der prädikatenlogischen Resolution kann eine logische Schlußfolgerung explizit unter Mitnahme dieser Eigenschaften geprüft werden. Es geht um eine logische Konsequenz, die mit Hilfe syntaktischer Verfahren abgebildet werden soll.

Es sei angenommen, daß - wie in Zeile 3 und 4 von Abb. 10 notiert - mit der analogen Quellseite (Originalseite) ein Problem besteht, jedoch ein Digitalisat vorhanden ist. Es sei hierfür ein terminologisches Wissen modelliert, das da ist:

- "digi" ist ein Digitalisat irgendeiner Originalseite.
- "sourceOf" ist eine Relation zwischen Digitalisat und Original. Jedes Digitalisat hat eine analoge Quelle.
- "digiproblem" ist eine Rolle für ein Digitalisat, wenn dessen Quelle nicht zur Gesamtquelle (Band), d.h. nicht zum Kontext passend ist.

- "correct" bedeutet, als Negation $\neg\,\mathrm{correct}$ gedacht: Die Quelle ist falsch, defekt, technisch zu schlecht gescannt oder nur ein Platzhalter (Dummy), der wie eine (bewußt) falsche Quelle betrachtet werden kann.

Als prädikatenlogische Formel ausgedrückt, sei dieses terminologische Wissen wie folgt ausgedrückt:

$$(\forall X)\ (digi(X)\ \rightarrow\ (\exists Y)\ sourceOf\,(Y,X))$$
$$(\forall X)\ (digiproblem(X) \leftrightarrow digi(X) \wedge \neg(\exists Y)[sourceOf\,(Y,X) \wedge correct(Y)])$$

Die erste Zeile drückt aus, daß für alle Dinge X gilt, wenn man eine Schnittmenge von allen X bildet, zu der dann nur die "digi"-Objekte gehören (das sind unsere erstellten Digitalisate), dann folgt daraus, daß es ein Y gibt, das die Quelle von X ist. Die zweite Zeile drückt aus: Ein "digiproblem" ist genau dann, wenn es ein "digi" ist und es kommt von keiner Quelle (Originalvorlage), die korrekt ist. Es liegt nun neben diesem terminologischen Wissen (T-Box) beispielsweise noch zusätzlich folgendes assertionales Wissen (A-Box) vor:

$digiproblem$(digipage110)
$sourceOf$ (analogpage120,digipage110)

Es ist nun die Frage, ob man den nachfolgenden Terminus logisch (also maschinell) schlußfolgern kann, daß nämlich Originalseite 120 irgendwie nicht korrekt *(correct)* ist, was wie folgt ausgedrückt wird: $\neg correct$(analogpage120)

Dafür wird die nun komplette Wissensbasis (in Abwandlung des Beispiels von [Sack11b]) als Formel notiert[31]:

((∀X)(digi(X)→(∃Y)sourceOf(Y,X))
∧(∀X)(digiproblem(X)↔
(digi(X) ∧¬(∃Y)(sourceOf(Y,X)∧correct(Y))))
∧ digiproblem(digipage110)
∧ sourceOf(analogpage120,digipage110))
→¬correct(analogpage120))

[31] Die Klammerfärbung ist für die leichtere Nachverfolgung der Terme vorgenommen.

Wenn diese Implikation[32] wahr ist, kann man die Aussage als allgemeingültig bezeichnen und damit ist tatsächlich das Objekt "analogpage120" nicht korrekt. Um die Allgemeingültigkeit zu zeigen, wird vor der Anwendung syntaktischer Umformungsregeln im Sinne eines Kalküls die Aufgabenstellung dahingehend festgelegt, daß die Unerfüllbarkeit der oben gegebenen Formelmenge nachgewiesen werden soll. Von der Unerfüllbarkeitsannahme muß ausgegangen werden, damit in richtiger Reihenfolge der Umwandlungsschritte bis zur Klauselform korrekt gearbeitet wird [Sack11b , Folie 42] und die Widerlegungsvollständigkeit der Klausel genutzt werden kann. Nachfolgende Beispielformeln sind von der Vorlesung [Sack00] zur Veranschaulichung auf das hier vorliegende Beispiel abgewandelt übernommen.

Zum Beweis der Unerfüllbarkeit muß der Formelmenge zur Annahme der Unerfüllbarkeit eine Negation vorangestellt werden.

¬((∀X)(digi(X)→(∃Y)sourceOf(Y,X))
∧(∀X)(digiproblem(X)↔
(digi(X) ∧¬(∃Y)(sourceOf(Y,X)∧correct(Y))))
∧ digiproblem(digipage110)
∧ sourceOf(analogpage120,digipage110))
→¬correct(analogpage120))

Der Schritt der Negationsnormalform[33] sei hier übersprungen. Nachfolgend wird die Pränexnormalform gezeigt, wo alle Quantoren nach außen gebracht sind [Schö00 S. 60]:

$(\forall X)(\exists Y)(\forall X1)(\forall Y1)(\forall X2)(\exists Y2)$
$((\neg digi(X) \vee sourceOf(Y,X))$
$\wedge(\neg digiproblem(X1) \vee (digi(X1) \wedge (\neg sourceOf(Y1,X1) \vee \neg correct(Y1)))$
$\wedge(digiproblem(X2) \vee (\neg digi(X2) \vee (sourceOf(Y2,X2) \wedge correct(Y2)))$
$\wedge$ *digiproblem*(*digipage*110)
$\wedge$ *sourceOf*(analogpage120,digipage110)
$\wedge$ *correct*(analogpage120))

[32] Im Prämissenteil der Implikation ist noch eine Implikation in der ersten Zeile enthalten. Gemeint ist aber das letzte Implikationszeichen.

[33] Normalformen sind z.B. „1“ von folgenden Formen {1, 01, 00001, 001} oder „A“ von {¬¬A, ¬¬¬¬A, ¬¬¬¬¬¬A} intuitiv zu verstehen.

Danach wird die Form in eine konjunktive Normalform (KNF) gebracht. Damit liegt eine Konjunktion ($\wedge$) von Disjunktionen ($\vee$) vor, die eine Voraussetzung für den Resolutionsalgorithmus ist.

$(\neg digi(X) \vee sourceOf(f(X),X))$
$\wedge(\neg digiproblem(X1) \vee digi(X1))$
$\wedge(\neg digiproblem(X1) \vee \neg sourceOf(Y1,X1) \vee \neg correct(Y1))$
$\wedge(digiproblem(X2) \vee \neg digi(X2) \vee sourceOf(g(X,X1,Y1,X2),X2))$
$\wedge(digiproblem(X2) \vee \neg digi(X2) \vee correct(g(X,X1,Y1,X2)))$
$\wedge$ digiproblem(digipage110)
$\wedge$ sourceOf(analogpage120,digipage110)
$\wedge$ correct(analogpage120)

Die letztendliche Klauselform sieht wie folgt aus:

1) {{¬digi(X),(sourceOf(f(X),X)},
2) {¬digiproblem(X1) , digi(X1)},
3) {¬digiproblem(X1), **¬sourceOf(Y1,X1)**, ¬correct(Y1)},
4) {digiproblem(X2), ¬digi(X2), sourceOf(g(X,X1,Y1,X2),X2)},
5) {digiproblem(X2), ¬digi(X2), correct(g(X,X1,Y1,X2))},
6) {digiproblem(digipage110)},
7) **{sourceOf(analogpage120,digipage110)**},
8) {correct(analogpage120)}}

Aus den 8 Zeilen werden nun die Klauseln herausgesucht, die resolviert werden können. Die Resolvente wird als zusätzliche Zeile notiert. Es können die Variablen *X* und *Y* aus dem terminologischen Wissen in Zeile 3 *¬sourceOf(Y1,X1)* durch die Konstanten aus dem assertionalen Wissen in Zeile 7 *analogpage120* und *digipage110* durch Unifizierung ersetzt werden. Somit wird ein Widerspruch zwischen

[sourceOf(analogpage120,digipage110)] und *¬[sourceOf(analogpage120,digipage110)]* sichtbar. Die Resolvente aus Zeile 3 ist damit *{¬digiproblem(X1), ¬correct(Y1)}* und wird als abgeleitete Klausel in Zeile 9 notiert:

1) {{¬digi(X),(sourceOf(f(X),X)},
2) {¬digiproblem(X1) , digi(X1)},
3) ~~{¬digiproblem(X1), **¬sourceOf(Y1,X1)**, ¬correct(Y1)},~~
4) {digiproblem(X2), ¬digi(X2), sourceOf(g(X,X1,Y1,X2),X2)},
5) {digiproblem(X2), ¬digi(X2), correct(g(X,X1,Y1,X2))},
6) {digiproblem(digipage110)},
7) ~~**{sourceOf(analogpage120,digipage110)**},~~
8) {correct(analogpage120)}}
9) {¬digiproblem(digipage110), ¬correct(analogpage120)}

Bei den Klauseln in Zeilen 8 und 9 sind ebenfalls einander widersprechende Klauselelemente mit Konstante *analogpage120* vorhanden. Die abgeleitete Klausel ist nun das Element von Zeile 9 *¬digiproblem(digipage110)* und diese neue Klausel kann –

mit Zeile 6 im Widerspruch stehend – zu der letzten Resolventen kommen, was geschlußfolgert die leere Klausel ist. An dieser Stelle stoppt der Algorithmus. Es ist damit ein Widerspruch für die negierte Wissensbasis abgeleitet und bewiesen, daß die Schlußfolgerung ¬*correct(analogpage120)* aufgrund der faktischen Wissensbasis erfüllend ist. Wäre die Schlußfolgerung jedoch in der ursprünglichen Wissensbasis wirklich falsch gewesen, wäre aufgrund der Nichtentscheidbarkeit der Resolution der Algorithmus ohne Haltepunkt vorangeschritten – er würde nie zu einem Ende führen.

Dieses einfache Beispiel sollte verdeutlichen, wie mit formallogisch formuliertem terminologischen und assertionalen Wissen auf der Basis einer modelltheoretischen Semantik eine Berechnung von implizitem Wissen vollzogen werden kann. Die komplexe formale und ausdrucksstarke Grundlage der Prädikatenlogik liefert keine Entscheidungssicherheit und RDF/RDFS ist semantisch jedoch nicht stark genug, weil z.B. keine Negation auszudrücken ist. Um eine Lokalität globaler Properties und Disjunktkeit von Klassen und eine semantische Begrenzung der Beziehung Subklassen zu Instanzen der Oberklasse, sowie Kardinalitätsrestriktionen (wie z.B.: ein Digitalisat hat genau eine analoge Seite als Quelle und nicht beliebig viele) und andere logischen Beziehungen ausdrücken zu können, bedarf es einer Beschreibungslogik, die zwar ausdrucksschwächer als die Prädikatenlogik ist, dafür aber die Entscheidbarkeit der Aussagenlogik aufrecht erhalten kann.

6. Goobi-Seiten als beschreibungslogisch geimpfte Objekte

Die Segmentierungen der XML-Strukturen, die in den Diagrammen vom Kernknoten ausgehend (Dg{1,2,3} E5-6) zu sehen sind (blaue Pfeile) sind als eine Abbildungstechnik von Überlappungen im XML-Dokumentenmodell eines streng mathematischen Baums vorhanden. Die Zuordnung der Seitenobjekte zu den Dokumentstrukturen ist in den Produktionsdaten und im validen METS über das Element <structLink> (siehe Abb. 19) über Referenzierungen gemacht. Eine lückenhafte Zuweisung von Seitenobjekten wie z.B. Paginierungen 85, 99, 100, 101, 102, 118 ist zwar in der Produktionsumgebung möglich, diese ist aber im Viewer nicht nachvollziehbar.

```
xlink:from="LOG_0005" xlink:to="PHYS_0001"/>
xlink:from="LOG_0005" xlink:to="PHYS_0015"/>
xlink:from="LOG_0005" xlink:to="PHYS_0016"/>
xlink:from="LOG_0005" xlink:to="PHYS_0017"/>
xlink:from="LOG_0005" xlink:to="PHYS_0018"/>
xlink:from="LOG_0005" xlink:to="PHYS_0035"/>
```

Abb. 25: Springende Bildfolge im Strukturelement (Ausschnitt aus <mets:structLink>)

Würde man im Viewer die erste Seite des Kapitels (das ist LOG_0005) aufschlagen, so würde der „page turner“ von PHYS_0001 als Landezone ausgehend nicht zu PHYS_0015 springen, sondern streng der Sortierung über das ORDER-Attribut [34] folgend auf PHYS_0002 zeigen, wenn man im Viewer die Blätterfunktion bedient. Die in Abb. 25 gezeigte Überlappung könnte hypothetisch eine andere Dokument(teil)sicht, z.B. das Zusammentragen von Abbildungen eines Künstlers, sein. Im METS-Standard (aber nicht bei Goobi) wird das Zusammenstellen von beliebigen Seiten zu einer „SEQUENCE OF FILES“ [Digi10 , S. 69/70] als Möglichkeit dargestellt, um jede beliebige Objektsequenz in die Navigationssequenz zu übersetzen.

[34] Das ORDER-Attribut befindet sich im XML-Element <mets:structMap TYPE="PHYSICAL"><mets:div TYPE=“BoundBook”…><<mets:div TYPE=”page”…OR DER=”1”>. Siehe in den Abb. 13, Abb. 14, Abb. 17 und Abb. 18.

Wege der Kooperation

Es gibt zwei Verfahrenswege, um dem Komplexitätszuwachs bei der Formulierung verschiedener semantischer Ebenen mit der Markupsprache XML auszuweichen. Auf der einen Seite ist das der Weg, komplett mit den Daten in die Sprachen des Semantic Web zu wechseln und zum anderen ist das die Methode einer semantischen Stand-Off-Annotation mit strikter Nutzung W3C-konformer Techniken des Semantic Web. Den Bezug auf das System von Goobi nehmend, ist die Wahl für die Kooperation aus folgenden Gründen für die nachfolgenden Erläuterungen getroffen. Die Größe des mittlerweile international platzierten Systems erlaubt keinen Austausch der Techniken. Die konventionelle Stabilität der METS-Dokumente ist groß. Sie verändern sich außerdem an keiner Stelle in einer Art, daß neue und auch noch extrem viel kumulierende Überlappungen hinzu kämen, weil die üblichen Benutzerschnittstellen das vorerst gar nicht zulassen. Als Transportbehälter sind sie nicht als erweiterte Annotationsgrundlage gedacht. Für dieses „unbestellte Anbaugebiet“ wäre jetzt der ideale Zeitpunkt gegeben, mit den Techniken des Semantic Web terminologisches und assertionales Wissen maschinell schlußfolgerungsfähig an diese Objekte von außen anzubinden. Auch die in den Goobi-Workflow integrierbare Volltextgenerierung mit einem Massen-OCR-Verfahren beliefert zur Zeit über einen quasi abgeschotteten Kanal die Indexer der Präsentationssysteme lediglich mit Rohdaten, um stichwortbasierte Volltextsuche zu ermöglichen. Die zukünftigen Erwartungen an Goobi könnten sich weiterhin auf die Transportfunktion der dokumentarischen Kernfracht konzentrieren, wobei die neuen „Frachtpapiere“ durch semantische Annotationswerkzeuge geschaffen, angereichert und auch in einen Kollaborationsmodus im Sinne einer Forschungsumgebung gebracht werden können.

Es soll mit dem Projekt EARMARK (Extremely Annotational RDF Markup) [Pero00a] die Kooperationsstrategie für das Diskursobjekt „Seite“ und dem Konzept „Digiproblem“ erörtert werden. EARMARK arbeitet mit RDF-Aussagen, wobei die nach dem Prinzip der Stand-off-Annotation konzipierten EARMARK-Dokumente in „OWL 2“-Ontologien modelliert sind. „EARMARK is based on an ontologically precise definition of markup that instantiates the markup of a text document as an independent OWL document outside of the text strings it annotates, and through appropriate OWL and SWRL characterizations it can define structures such as trees or graphs and can be used to generate validity constraints (including co-constraints currently unavailable in most

validation languages).” [Pero00a] Über das OWL-Dokument können Abfragesprachen wie SPARQL [GaSe00] und Semantic Reasoner wie „Pellet: OWL 2 Reasoner for Java“ [00ad] oder „Apache Jena“ [00ae] genutzt werden.

Die Seitenobjekte PHYS_(0001 … 9999) (Dg3 BJ9-10 bzw. K4-9) sind syntaktisch voneinander unabhängige Objekte, die in sich überlappenden Konzeptstrukuren liegen. Diese Entitäten sind aber ein geeigneter Ansatzpunkt, um sie im terminologischen Wissensbereich mit dem „Digiproblem“-Konzept in Form von Kollektionen oder Listen zu verbinden. Wenn sie als Einzelobjekte zudem konzeptionell identisch sind[35], dann können explizit semantische Veränderungen über beschreibungslogische Mechanismen in eine Äquivalenzklasse eingeführt und wieder herausgeführt werden. Dafür ist das syntaktische „OWL 2 DL“-Element „hasKey“ [36] geeignet. Der „semantische Kippschalter“ der Property „correct“ (siehe Seite 61) kann je nach Klientel und Abfragemotivation, den mit Bildfehlern behafteten Band als Ganzes oder nur in Teilen „stigmatisieren“. Besteht z.B. das Interesse, eine Seite aus einem Band zu referenzieren, die nicht im problembehafteten Container „Kapitel 5“ enthalten ist, dann unterscheidet sich diese Nutzung von der eines Exports des ganzen Bandes, wo eine problematische Quellensituation durchaus relevant ist.

Die Autoren des EARMARK-Projektes argumentieren in [IoPV11] S. 1706, daß RDF noch nicht das Vokabular besitzt, um die Ressourcen in deren semantisch expliziten Beziehungen zu fassen. Sie können zwar mit RDF aus dem strukturellen Markup in XML als Elemente, Attribute, Kommentare und Text identifiziert werden, würden jedoch in der Ausdruckskraft von RDF verbleiben. Stärkere Ausdrücke könnte wiederum das Vokabluar von RDFS bzw. eines der OWL-Sprachen bieten. Die Autoren raten jedoch ab, Beschreibungssprachen (dort RDF und OWL 2 DL) zu mischen und weisen auf bekannte praktische Grenzen hin, die eine solche Kombination besitzt. Um einen Ersatz für das eigentlich gewünschte RDF-Modell von Containern zu erhalten, wird auf eine in die EARMARK-Ontologie [PeIV11] (siehe Anhang) einbezogene „Collections Ontology“ [CiPe00] von [Cicc08a] [Cicc08b] – aufbauend auf [DRSM06] – hingewiesen. Der

[35] Die lokale ID der PHYS_xxxx-Objekte läßt darauf schließen

[36] owl:hasKey - OWL Test Cases. URL http://owl.semanticweb.org/page/Owl:hasKey. - abgerufen am 20.05.2013 und am 14.04.2014 mit Ladefehler.

Inkubator für die „Collections Ontology“ [CiPe00] ist die „List Ontology“ von [DRSM06] und erstere ist wiederum in die EARMARK-Ontologie integriert.

Es wird ebenso vermerkt, daß das RDF-Vokabular - neben seiner semantischen Schwäche von rdf:List, rdf:Alt, rdf:Bag, rdf:Seq - ohnehin nicht in OWL DL verwendet werden kann, weil RDF in der OWL-Serialisierung benutzt wird. Zur erlaubten Alternative rdf:Seq wird ausgesagt: „Although rdf:Seq is not illegal, it depends on lexical ordering and has no logical semantics accessible to a DL classifier.” [DRSM06] Diese Einschränkung erinnert an die strikte ganzzahlig aufsteigende Sortierung der PHYS_xxxx-Objekte in Goobi, die – wie in Kapitel 4.4 vorgestellt – eine Sortierung methodisch an die Semantik des Bildstapels fesselt und nicht die Bedeutungen des Inhalts freigibt. In der „Collections Ontology“ gibt es dafür die OWL-Klassen „List“, „Bag“, „Set“, „Collection“ als frei relativierbare Behälter mit Sinn.

METS-Dokumente und auch das Goobi-Produktionsdatenformat in XML sind bestückt mit Sequenzen, die auftragsgemäß die sequentielle Struktur der Druckwerke in das digitale Derivat migriert haben. Dieses Derivat bedient nach den DFG-Richtlinien die METS-Anwendungsklasse der „page turner“. Es handelt sich hier um topologische Sortierungen verschiedener Art, die im XML-Dokumentenmodell mit relativ viel Aufwand codiert werden müssen. Die Topologien können formal-semantisch mit OWL 2 DL, SWRL [37] und dem EARMARK-Dokumentenmodell differenzierter aufgegriffen werden.

Mit Mitteln aus dem terminologischen Wissen kann ein interpretierter Abschnitt der METS-Datei über eine Klasseninstanz *docuverse* von der Klasse *Docuverse* von einer Klasseninstanz *range* her referenziert werden. Die Klasseninstanz *range* umfaßt – weil hier ein XML-Dokument vorliegt – einen Abschnitt der METS-Datei, der mit einem in der xpath-Sprache formulierten Ausdruck beschrieben wird. Dieser Ausdruck ist der Wert der Property *has xpath context* und weist auf eine Subklasse von der Klasse *Range*, nämlich die Klasse *xpath range*. Über die „data property“ *has xpath context* kann über

[37] SWRL (Semantic Web Rule Language • http://www.w3.org/Submission/SWRL/) übernimmt für das EARMARK Framework Regelbeschreibungen, die OWL nicht abdeckt. “OWL does not allow direct pointer arithmetics, SWRL on the contrary does, as shown earlier where we described a batch of (SWRL-implementable) rules that do, in fact, determine overlapping locations on EARMARK documents with good efficiency.” [IoPV11]

eine Klasseninstanz von *xpath pointer range* ein Zeichenbereich definiert werden, der somit keine Einzelzeichenlokalisierung beim dem Setzen Anfangs- und Endmarken des Bereichs benötigt. „EARMARK allows references to relatively stable fragment ids of the documents (by using XPath ranges without specifying explicitly begin and end locations) rather than the extremely fragile character locations, further reducing the chances of outdated pointers." [IoPV11 , S. 1706] In [PeVi09] wird noch bedauert, daß das XPointer Schema [GMMD03] zu keinem W3C-Standard gemacht wurde, mit dem Inhalt in XML-Dateien adressierbar wäre.

Die *ranges* müssen nicht mehr hierarchisch organisiert werden, sondern können beliebig sequenziell geschrieben und können sich beliebig einander überlappen. Damit ist eine flache Repräsentation eines EARMARK-Dokuments vorgestellt und anwendbar, wo es um Texte, wie zum Beispiel Volltextkorpora von OCR-Prozessen herrührend geht. Um das markup der METS-Dokumente in ein EARMARK-Dokument zu transferieren stehen Automatismen zur Verfügung. [PeVi09 , Kap. 4.1] Das Element- und Attribut-Mapping ermöglicht damit auch, die mit URIs referenzierbaren Daten aus dem MODS-Teil des METS-XML als Normdatenverknüpfungen für das Semantic Web sichtbar zu machen. Lokale Daten, wie Pixelauflösung, MIME-Type, Dateipfad oder lokale „digiproblem"-Markierungen könnten aus Bestandsdaten und Neudaten gemixt werden. Die Transformation von XML zu RDF/XML von EARMARK ist verlustlos in beide Richtungen. Es gibt einen Demonstrator unter [Pero00b], der XML-Quellen in eine OWL-Ontologie umwandelt, die mit der EARMARK-Ontologie konform ist. Bei EARMARK ist die Idee, wie die Autoren der Ontologie angeben, von Ted Nelsons „Project Xanadu" [Nels00] aufgegriffen worden.

Mit diesem Sprachkonstrukt könnten somit beliebige METS-Dokumentsequenzen mit den schon genannten Schlußfolgerungswerkzeugen (SPARQL, Pellet, Jena – siehe Seite 68) abgefragt bzw. bestimmt werden. Es könnte damit eine Sequenz von Seiten geprüft werden, ob sich darin eine „digiproblem"-markierte Seite[38] befindet und zur welcher übergeordneten Sequenz diese Seite oder Sequenz zugeordnet ist. [DRSM06] nutzt das Modellieren von Wissensbasen zum logikgestützten Rechnen mit Aminosäuresequenzen in der Eiweißforschung. Ein isolierbarer Teil des Goobi-Diskursuniversums mit einer

[38] siehe zu „digiproblem" Seite 61

Ausdehnung auf 80 Stukturelementtypen (siehe Seite 33) und den Seitenelementen ist mit der EARMARK-Ontologie ähnlich ideal interpretierbar. Anhand dieser kleinen Facette des Dokumentenmanagements soll gezeigt worden sein, daß das Problem fehlender oder falscher Originalseiten nicht verschoben werden, bzw. als zu selten auftretend hingenommen werden muß. Für die Seltenheitsbehauptung gibt es keinen Beweis bzw. kein mögliches maschinelles Beweisverfahren, um diese Behauptung zu stützen.

7. Ausblick

In der Arbeit wurde die Strukturanalogie von Goobi-Digitalisaten in Bezug auf deren gedruckten Quellen um das Modell einer formalsemantisch beschriebenen Digitalobjekt-Originalseite-Beziehung erweitert und für mindestens einen wichtigen Punkt ein kontrollillusorischer Qualifikationsüberschuß festgestellt (Seite 22). Die nach außen hin dokumentarisch auftretenden „page turner im nur-lesen-Modus“ DFG-Viewer & Co. bekommen ihr Material von Produktionssystemen wie z.B. Goobi oder Visual Library[39]. Diese Systeme befinden sich – von den förderregulierenden Körperschaften gestützt und von den produzierenden Einrichtungen nach wie vor als Seitenerntemaschinen mit Schaufenster zusammengeschaltet – noch mitten in einer nationalen Auftragsserie. Das Einzelprodukt der Massendigitalisierung durch angelernte Hilfskräfte oder durch Fachabteilungen erstellt, mit den Güteklassen „bitonal, kaum lesbar“ bis zu einer Editionsvorstufe (Seite 15), beschäftigt sich noch nicht mit einer effizienten Beantwortung der Vollständigkeitsfrage auf Werkebene. Die regulativen Schwachpunkte bei dem Einbezug von URIs sind ebenfalls dem Transportvorrang zum Datenspeicher hin geschuldet.

Das Datenmodell von Goobi ist von einer Strukturanalogie geprägt, die wiederum selbst aus dem Analogieziel Archivierung und Authentizität hervorgeht. Das elektronische Faksimile kann durch dieses Datenmodell METS abgebildet werden. Besteht jedoch der Wunsch, mit automatischen Schlußfolgerungen mehr Wissen aus vorhandenem Wissen ableiten zu können, so muß über diese Faksimile-Ebene eine explizit beschriebene semantische Ebene gebracht werden. Ein Beispiel für diese Wissensgenerierung wurde an einem Fehlermanagement für verlustige Originalseiten aufgezeigt. Als Konterpart für dieses Fehlermanagement muß, wie in Kapitel 4.3.- 4.5 analysiert wurde, unbedingt das Programm für Bildstapelkorrekturen verändert werden, um die URN-Bindungen an die Seitenelement nicht zu verlieren.

[39] http://www.walternagel.de/visual-library

Eine Öffnung von Goobi für Techniken des Semantic Web benötigt Personen, die dieses umsetzen möchten. Einem Start mit Goobi geht für die meisten Interessenten eine Lernphase voraus, weil das System keine out-of-the-box-Anwendung ist. Die verstreuten Dokumente um das Datenmodell und die Architektur von Goobi sollte in dieser Arbeit bei weitem nicht vollständig geordnet wiedergegeben sein. Hinzu kommt das hier nicht behandelte verstreute Wissen der Goobi-Anwender in den Einrichtungen, das zusammengebracht sehr hilfreich ist.

Es sollte ein interdisziplinäres Team die Möglichkeiten diskutieren, ob diese Java-Anwendung [40] EARMARK für Goobi den erstmaligen Anschluß an Semantic Web Technologien bildet. Es wäre von großem Vorteil, wenn man mit Hilfe von Schlußfolgerungsmechanismen aus dem sehr großen gemeinsamen Datenbestand neues Wissen schöpfen könnte. Hier seien nur einige Beispielfragen vorgestellt:

Es sind in Goobi-Projekten bisweilen Platzhalterbilder für Lücken im Quellmaterial in den Bildstapel eingesetzt, jedoch nicht in den Metadaten, wie z.B. in den Inhaltsverzeichnissen (aus verständlichen Gründen) notiert. Wie können diese Bilddateien wieder gefunden werden? Anhand welchen assertionalen Wissens, können solche Platzhalterlokalisierungen geschlußfolgert werden? Wie könnte man eine automatische Suche nach Ersatzmaterial für als lückenhaft markierte Digitalisate in Bibliothekskatalogen gestalten? Sollte eine Volltextsuche, die bisher über einen Gesamtindex pro Viewer, auch in einem eingeschränkten Materialbereich, wie z.B. Zeitschrift, Zeitschriftenjahrgang, Band, Heft, Artikel nicht sinnvoll sein? Wären individuell gestaltbare topologische Sichten, wie z.B. Abbildungskollektionen oder Sammlungen von Paratexten [Gene89] gewünschte Materialsichten?

Je mehr Erfahrung bei dem Einsatz der Semantic Web Technologie zur Wissensvermittlung in den „Frachtpapieren“ gesammelt werden, desto „ruhiger“ kann die Fracht in den METS-Dateien lagern, denn letztlich sind METS-Dateien mit ihrem schmalen Aufgabenheft nicht viel anderes als die Dateisysteme mit den zugehörigen

[40] „Dependencies:
* OWLAPI 2.2.0 (http://downloads.sourceforge.net/owlapi/owlapi-2.2.0.zip?use_mirror=osdn)
* Jena 2.6.0 (http://prdownloads.sourceforge.net/jena/jena-2.6.0.zip?download)
* Pellet 2.0 (http://clarkparsia.com/pellet/download/pellet-2.0.0)“
https://github.com/essepuntato/EarmarkDataStructure / - abgerufen 14.04.2014

Digitalisaten – beide funktionieren gut als strukturierte Lager. Über die „Lagersprache“ der digitalen Klone von Druckwerken hinausgehende Aussagen gehören aber in Sprachen geschrieben, deren formal definierte Semantik Entscheidbarkeit, Vollständigkeit und Korrektheit bietet. Bei dem momentanen Nutzungsstand des Goobi-Dokumentenmodells wäre kein Exodus aus METS nötig, weil noch nichts drin ist. Man könnte vielmehr neu beginnen.

8. Quellennachweis

Die Literaturverwaltung wurde mit Zotero 4.0.8 i.V.m. Mircosoft Word und der Firefox-Erweiterung „Zotero Word for Windows Integrator 3.1.12“ gesteuert. Der Zitationsstil ist nach DIN 1505-2 (alphanumerisch, deutsch).

[00a] *Category:Reasoner - Semantic Web Standards*. URL http://www.w3.org/2001/sw/wiki/Category:Reasoner. - abgerufen am 2014-04-14

[00b] *Goobi Community Edition in Launchpad*. URL https://github.com/goobi. - abgerufen am 2014-04-14

[00c] *Goobi - Digital Library Modules: Goobi.Production*. URL http://www.goobi.org/software/goobiproduction/. - abgerufen am 2014-04-14

[00d] *Goobi - Digital Library Modules: Goobi.Presentation*. URL http://www.goobi.org/software/goobipresentation/. - abgerufen am 2014-04-14

[00e] *intranda viewer*. URL http://www.digiverso.com/de/products/viewer. - abgerufen am 2014-04-14

[00f] *Goobi - Digital Library Modules: Anwender*. URL http://www.goobi.org/community/anwender/. - abgerufen am 2014-04-14

[00g] *DFG - Deutsche Forschungsgemeinschaft - DFG-Praxisregeln „Digitalisierung“*. URL http://www.dfg.de/formulare/12_151/. - abgerufen am 2014-04-14

[00h] *DFG - Deutsche Forschungsgemeinschaft - Veröffentlichungen*. URL http://www.dfg.de/foerderung/programme/infrastruktur/lis/veroeffentlichungen/index.html#micro10402394. - abgerufen am 2014-04-14

[00i] *Guidelines: Technical Guidelines for Digitizing Cultural Heritage Materials - Federal Agencies Digitization Guidelines Initiative*. URL http://www.digitizationguidelines.gov/guidelines/digitize-technical.html. - abgerufen am 2014-04-14

[00j] *Werkansicht | Digitalisierte Sammlungen der SBB. „Christiani Vita Et Corona“*. URL http://digital.staatsbibliothek-berlin.de/dms/werkansicht/?PPN=PPN651724848&LOGID=LOG_0005. - abgerufen am 2014-04-24

[00k] *MODS: Uses and Features (Metadata Object Description Schema: MODS)*. URL http://www.loc.gov/standards/mods/mods-overview.html

[00l] *Was sind und was sollen Bibliothekarische Datenformate*. URL http://www.allegro-c.de/formate/formate.htm. - abgerufen am 2014-04-14

[00m] *Digitalisierungsmetadaten - dini-ag-kim - Deutsche Nationalbibliothek - Wiki*. URL https://wiki.dnb.de/display/DINIAGKIM/Digitalisierungsmetadaten. - abgerufen am 2014-04-14

[00n] *DFG-Viewer: Über das Projekt*. URL http://dfg-viewer.de/ueber-das-projekt/. - abgerufen am 2014-04-14

[00o] *John Unsworth*. URL http://people.brandeis.edu/~unsworth/. - abgerufen am 2014-04-14

[00p] *DFG - Deutsche Forschungsgemeinschaft - Evaluationsstudien*. URL http://www.dfg.de/dfg_profil/foerderatlas_evaluation_statistik/programm_evaluation/studien/index.html. - abgerufen am 2014-04-14

[00q] *Uniform Resource Name – Wikipedia*. URL http://de.wikipedia.org/w/index.php?title=Uniform_Resource_Name&oldid=117226912. - abgerufen am 2013-05-01

[00r] *IANA | URI Schemes*. URL http://www.iana.org/assignments/uri-schemes.html. - abgerufen am 2014-04-14

[00s] *Uniform Resource Names (URN) Namespaces*. URL http://www.iana.org/assignments/urn-namespaces/urn-namespaces.xml. - abgerufen am 2014-04-14

[00t] *NBN-Pruefziffer-Berechnung*. URL http://nbn-resolving.de/nbnpruefziffer.php. - abgerufen am 2014-04-14

[00u] *Persistent Identifier - Systembeispiele*. URL http://www.persistent-identifier.de/ueberblick/Beispiele.php. - abgerufen am 2014-04-14

[00v] *DNB / Standardisierung / Persistent Identifier*. URL http://www.dnb.de/DE/Standardisierung/PI/pi_node.html. - abgerufen am 2014-04-14

[00w] *DNB - URN-Service*. URL http://www.dnb.de/urnservice.html. - abgerufen am 2014-04-14

[00x] *METS Example Documents: Metadata Encoding and Transmission Standard (METS) OfficialWeb Site*. URL http://www.loc.gov/standards/mets/mets-examples.html. - abgerufen am 2014-04-14

[00y] *Abschnitt: Wer entwickelt Goobi?* URL http://www.digiverso.com/de/products/goobi. - abgerufen am 2014-04-14. — Goobi Workflow. Populäre Open-Source-Software

[00z] *ULB Sachsen-Anhalt - Digitale Sammlungen des 16., 17. und 18. Jahrhunderts*. URL http://digitale.bibliothek.uni-halle.de. - abgerufen am 2014-04-14

[00aa] *Digitale Kollektionen der SLUB Dresden*. URL http://digital.slub-dresden.de/kollektionen/. - abgerufen am 2014-04-14

[00ab] *SLUB Dresden: Werkansicht: Atlas selectus von allen Königreichen und Ländern der Welt*. URL http://digital.slub-dresden.de/werkansicht/cache.off?id=5363&tx_dlf[id]=1425&tx_dlf[page]=17. - abgerufen am 2014-04-14

[00ac] *20 Non-hierarchical Structures - TEI P5: — Guidelines for Electronic Text Encoding and Interchange*. URL http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html. - abgerufen am 2014-04-14

[00ad] *Pellet: OWL 2 Reasoner for Java*. URL http://clarkparsia.com/pellet. - abgerufen am 2014-04-14

[00ae] *Apache Jena - Apache Jena*. URL http://jena.apache.org/. - abgerufen am 2014.-04-14

[09] *VCC - Kompetenzzentrum für Videokonferenzdienste: News - Jahr 2002*. URL http://vcc.zih.tu-dresden.de/index.php?linkid=20000&jahr=2002. - abgerufen am 2014-04-14

[12] *Policy für die Vergabe von URNs im Namensraum urn:nbn:de [Elektronische Ressource] / Deutsche Nationalbibliothek. Uta Ackermann ; Christiane Berner ; Natalie Elbert ; Jürgen Kett ; Kadir Karaca Kocer ; Nicole von der Hude ; Martina Wiegand*. Version 1.0; Stand: 29. November 2012. Aufl. Frankfurt, M. : Deutsche Nationalbibliothek, 2012

[97] *URN Syntax / Request for Comments: 2141*. URL http://www.ietf.org/rfc/rfc2141. - abgerufen am 2014-04-14. — RFC 2141

[AlMa05] ALEXANDER CZMIEL ; MANFRED THALLER, 1950- [RED.] (Hrsg.): *Retrospektive Digitalisierung von Bibliotheksbeständen : Evaluierungsbericht über einen Förderschwerpunkt der DFG* : Köln : Univ., 2005 — ISBN Köln

[AnHa08] ANTONIOU, GRIGORIS ; HARMELEN, FRANK VAN: *A semantic Web primer*, *Cooperative information systems*. 2. ed. Aufl. Cambridge, Mass. [u.a.] : MIT Press, 2008 — ISBN 978-0-262-01242-3

[Bern94] BERNERS-LEE, TIM: *RFC 1630 - Universal Resource Identifiers in WWW: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web*. URL http://tools.ietf.org/html/rfc1630. - abgerufen am 2014-04-14

[Burd10] BURDETTE, ALAN: The EVIA Digital Archive Project: Challenges and Solutions (2010)

[Chen12] CHEN, ESTHER: Linked Open VD 17 — von METS/MODS zum Europeana Data Model. In: , *Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft - (Berliner Handreichungen)*. Bd. 327 : Humboldt-Universität zu Berlin, Philosophische Fakultät I, Institut für Bibliotheks- und Informationswissenschaft, 2012

[Chom70] CHOMSKY, NOAM: *Aspekte der Syntax-Theorie. Aspects of the theory of syntax <dt.>*. Berlin : Berlin : Akademie-Verl., 1970

[Chom75] CHOMSKY, NOAM: *Current issues in linguistic theory*. 6. print. Aufl. The Hague : de Gruyter, 1975 — ISBN 90-279-0700-5

[Cicc08a] CICCARESE, PAOLO: *HckLab - Working foR evolution: Moving towards the SWAN Collections Ontology [1]*. URL http://hcklab.blogspot.de/2008/12/moving-towards-swan-collections.html. - abgerufen am 2014-04-14

[Cicc08b] CICCARESE, PAOLO: *HckLab - Working foR evolution: Moving towards the SWAN Collections Ontology [2]*. URL http://hcklab.blogspot.de/2008/12/moving-towards-swan-collections_31.html. - abgerufen am 2014-04-14

[CiPe00] CICCARESE, PAOLO ; PERONI, SILVIO: *collections-ontology - Collections Ontology (CO) - Google Project Hosting*. URL http://code.google.com/p/collections-ontology/. - abgerufen am 2014-04-14

[Dero04] DEROSE, STEVEN: *Markup Overlap: A Review and a Horse*. URL http://conferences.idealliance.org/extreme/html/2004/DeRose01/EML2004DeRose01.html. - abgerufen am 2014-04-14. — Proceedings of Extreme Markup Languages®. — DeRose_Markup-Overlap_Review-and-a-Horse_2004.pdf

[Deut13] Anhang A: METS/MODS-Profil für die Darstellung im DFG-Viewer und Übermittlung per OAI. In: DEUTSCHE FORSCHUNGSGEMEINSCHAFT (DFG) (Hrsg.) (2013)

[Digi10] DIGITAL LIBRARY FEDERATION: *<METS> METADATA ENCODING AND TRANSMISSION STANDARD: PRIMER AND REFERENCE MANUAL. Version 1.6 Revised*. URL http://www.loc.gov/standards/mets/METSPrimerRevised.pdf. - abgerufen am 2014-04-14

[DRSM06] DRUMMOND, NICK ; RECTOR, ALAN L. ; STEVENS, ROBERT ; MOULTON, GEORGINA ; HORRIDGE, MATTHEW ; WANG, HAI ; SEIDENBERG, JULIAN: Putting OWL in Order: Patterns for Sequences in OWL. In: *OWLED*, 2006

[EnFu00a] ENDERS, MARKUS ; FUNK, STEFAN E.: *zvdd/DFG-Viewer METS-Profil, Version 2.0, 16.04.2009 (XML-Version)*. URL http://dfg-viewer.de/fileadmin/groups/dfgviewer/METS_Anwendungsprofil_2.0.xml. - abgerufen am 2014-04-14. — METS_Anwendungsprofil_2.0.xml

[EnFu00b] ENDERS, MARKUS ; FUNK, STEFAN E.: *UGH - Java Metadaten Bibliothek – Goobi*. URL http://wiki.goobi.org/index.php/UGH_-_Java_Metadaten_Bibliothek. - abgerufen am 2014-04-14

[Funk09] FUNK, STEFAN E.: zvdd/DFG-Viewer METS-Profil – Version 2.0 (PDF-Version) (2009)

[GaSe00] GARLIK, STEVE HARRIS ; SEABORNE, ANDY: *SPARQL 1.1 Query Language*. URL http://www.w3.org/TR/2013/REC-sparql11-query-20130321/. - abgerufen am 2014-04-14

[Gene89] GENETTE, GÉRARD: *Paratexte : das Buch vom Beiwerk des Buches*. 1. Aufl. Frankfurt/M. [u.a.] : Campus-Verl. [u.a.], 1989 — ISBN 3-593-34061-5

[GMMD03] GROSSO, PAUL ; MALER, EVE ; MARSH, JONATHAN ; DEROSE, STEVEN: *XPointer element() Scheme*. URL http://www.w3.org/TR/2003/REC-xptr-element-20030325/. - abgerufen am 2014-04-14

[Grub93] GRUBER, THOMAS S.: *Toward Principles for the Design of Ontologies Used for Knowledge Sharing (onto-design.pdf)*. URL http://tomgruber.org/writing/onto-design.htm. - abgerufen am 2014-04-14. — Home » Publications and Presentations » Toward Principles for the Design of Ontologies Used for Knowledge Sharing

[Hitz08] HITZLER, PASCAL: *Semantic Web : Grundlagen*. 1. Aufl. Aufl. Berlin ua : Springer, 2008 — ISBN 3-540-33993-0

[IoPV10] IORIO, ANGELO ; PERONI, SILVIO ; VITALI, FABIO: Handling Markup Overlaps Using OWL. In: CIMIANO, P. ; PINTO, H. S. (Hrsg.): *Knowledge Engineering and Management by the Masses*, *Lecture Notes in Computer Science*. Bd. 6317 : Springer Berlin Heidelberg, 2010 — ISBN 978-3-642-16437-8, S. 391–400

[IoPV11] DI IORIO, ANGELO ; PERONI, SILVIO ; VITALI, FABIO: A Semantic Web approach to everyday overlapping markup. In: *Journal of the American Society for Information Science and Technology* Bd. 62 (2011), Nr. 9, S. 1696–1716

[Nels00] NELSON, THEODOR HOLM (TED NELSON): *Project Xanadu®*. URL http://www.xanadu.com/. - abgerufen am 2014-04-14

[NOSS00] NEUROTH, HEIKE ; OßWALD, ACHIM ; SCHEFFEL, REGINE ; STRATHMANN, STEFAN ; HUTH, KARSTEN: *nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3*. v2.3 (Online.-Version). Aufl. — nestor ist ein vom BMBF gefördertes Projekt zum Aufbau eines Kompetenznetzwerkes zur Langzeitarchivierung digitaler Objekte in Deutschland. Die nestor Informationsdatenbank ist Bestandteil eines Subject Gateways zu Fragen der Langzeitarchivierung digitaler Objekte.

[PeIV11] PERONI, SILVIO ; DI IORIO, ANGELO ; VITALI, FABIO: *EARMARK Ontology - Current version 1.8.1*. URL

http://www.essepuntato.it/lode/imported/http://www.essepuntato.it/2008/12/earmark. - abgerufen am 2014-04-14

[Pero00a] PERONI, SILVIO: *EARMARK | Mark everything up!* URL http://palindrom.es/phd/research/earmark/. - abgerufen am 2014-04-14

[Pero00b] PERONI, SILVIO: *XML To EARMARK conversion tool*. URL http://speronitomcat.web.cs.unibo.it:8080/XML2EARMARK/. - abgerufen am 2014-04-14

[PeVi09] PERONI, SILVIO ; VITALI, FABIO: Annotations with EARMARK for arbitrary, overlapping and out-of order markup. In: *Proceedings of the 9th ACM symposium on Document engineering*, *DocEng '09*. New York, NY, USA : ACM, 2009 — ISBN 978-1-60558-575-8, S. 171–180

[Powe03] POWERS, SHELLEY: *Practical RDF*. 1. Aufl. : O'Reilly, 2003 — ISBN 978-0596002633

[Sack00] SACK, HARALD: *Semantic Web Technologien - Aussagenlogik und Prädikatenlogik - am Hasso Plattner Institut, Universität Potsdam - 5. Vorlesung*. URL http://yovisto.de/video/19462. - abgerufen am 2014-04-14

[Sack11a] SACK, HARALD: *Folien zur 4. Vorlesung „Semantic Web Technologien" am Hasso Plattner Institut, Universität Potsdam, Wintersemester 2011/12, am 15.11.2011 : Ontologien in Philosophie und Informatik*. URL http://de.slideshare.net/lysander07/04-ontologie-in-der-philosophie-und-der-informatik-semantic-web-technologien-ws-201112. - abgerufen am 2014-04-14

[Sack11b] SACK, HARALD: *Folien zur 5. Vorlesung „Semantic Web Technologien" am Hasso Plattner Institut, Universität Potsdam, Wintersemester 2011/12 am 22.11.2011 : Aussagenlogik und Prädikatenlogik*. URL http://de.slideshare.net/lysander07/05-aussagenlogik-und-prdikatenlogik-semantic-web-technologien-ws-201112. - abgerufen am 2014-04-14

[ScDA08] SCHOGER, ASTRID ; DOBRATZ, SUSANNE ; ALTENHÖNER, REINHARD: *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive - Version 2 -, Nestor-Materialien*. Bd. nestor-Kriterien. Frankfurt a. M. : nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland, 2008

[Schö00] SCHÖNING, UWE: *Logik für Informatiker*. Heidelberg [u.a.] : Spektrum Akad. Verl., 2000 — ISBN 3-8274-1005-3

[Sowa00] SOWA, JOHN F.: *Ontology*. URL http://www.jfsowa.com/ontology/. - abgerufen am 2014-04-14

[Unsw10] UNSWORTH, JOHN: *EVIA, Sustainability, and Mission-Creep*. URL http://cnx.org/content/m34341/latest/. - abgerufen am 2014-04-14. — Connexions

[Varg70] VARGA, TAMÁS: *Mathematische Logik für Anfänger. 1. Aussagenlogik. Matematikai logika, kezdöknek <dt.>*. Berlin : Volk und Wissen, 1970

[Vran10] VRANDECIC, ZDENKO: *Ontology Evaluation*. Karlsruhe, Fakultät für Wirtschaftswissenschaften (WIWI) Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB), 2010

[Wagn00a] WAGNER, KARL HEINZ: *Kapitel 3: Semiotik - Kap. 3.8 System und Struktur*. URL http://www.fb10.uni-bremen.de/khwagner/grundkurs1/kapitel3.aspx#System-und-Struktur. - abgerufen am 2014-04-14. — Einführung in die Sprachwissenschaft

[Wagn00b] WAGNER, KARL HEINZ: *Einführung in die Sprachwissenschaft. Kapitel 2: Allgemeine Grundbegriffe - Wissenschaft*. URL http://www.fb10.uni-bremen.de/khwagner/grundkurs1/kapitel2.aspx. - abgerufen am 2014-04-14

9. Anhang

Zur Veranschaulichung dienen die beigefügten Diagramme in diesem Anhang. Die für diese Untersuchung neu erstellten Diagramme haben als Orientierungshilfe eine Spalten- und Reihennotation an den Blatträndern. Orientierungshinweise werden im Haupttext geschrieben im Sinne von „Diagramm 1 Spalte E bis F Reihe 5 bis 6," und dies in Kurzform: (Dg1 EF5-6).

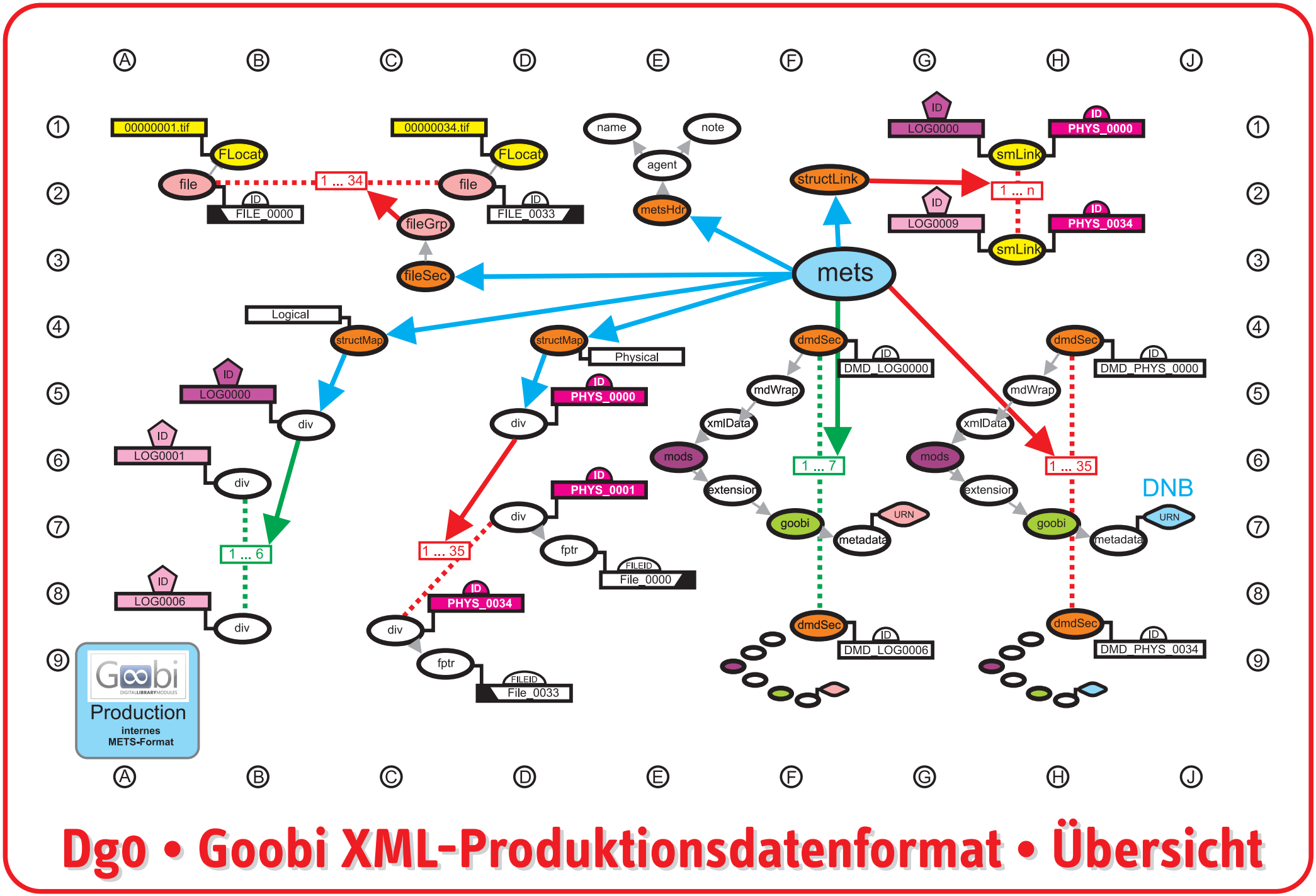
9.1 Erklärung zu Diagramm 0

Diagramm 0 zeigt das XML-Produktionsformat von Goobi. Es handelt sich nicht um valides METS, sondern um eine Vorform. Das Diagramm 0 veranschaulicht als Überblick (ohne alle Details) die Segmentierung. Die Intervallabbildung wie (Dg0 BC2) [1…34] soll bedeuten, daß sich die Teilstruktur ebensoviel wiederholt, wie angegeben. Die verkleinerten Strukturen in (Dg0 FJ9) sind nur aus Platzgründen gewählt. Im Diagramm 0 sind alle XML-Elemente, aber nicht alle Attribute dargestellt. Die Namensräume wurden bei den Elementnamen weggelassen. Diese sind in den ebenfalls angehangenen XML-Dokumenten zu sehen.

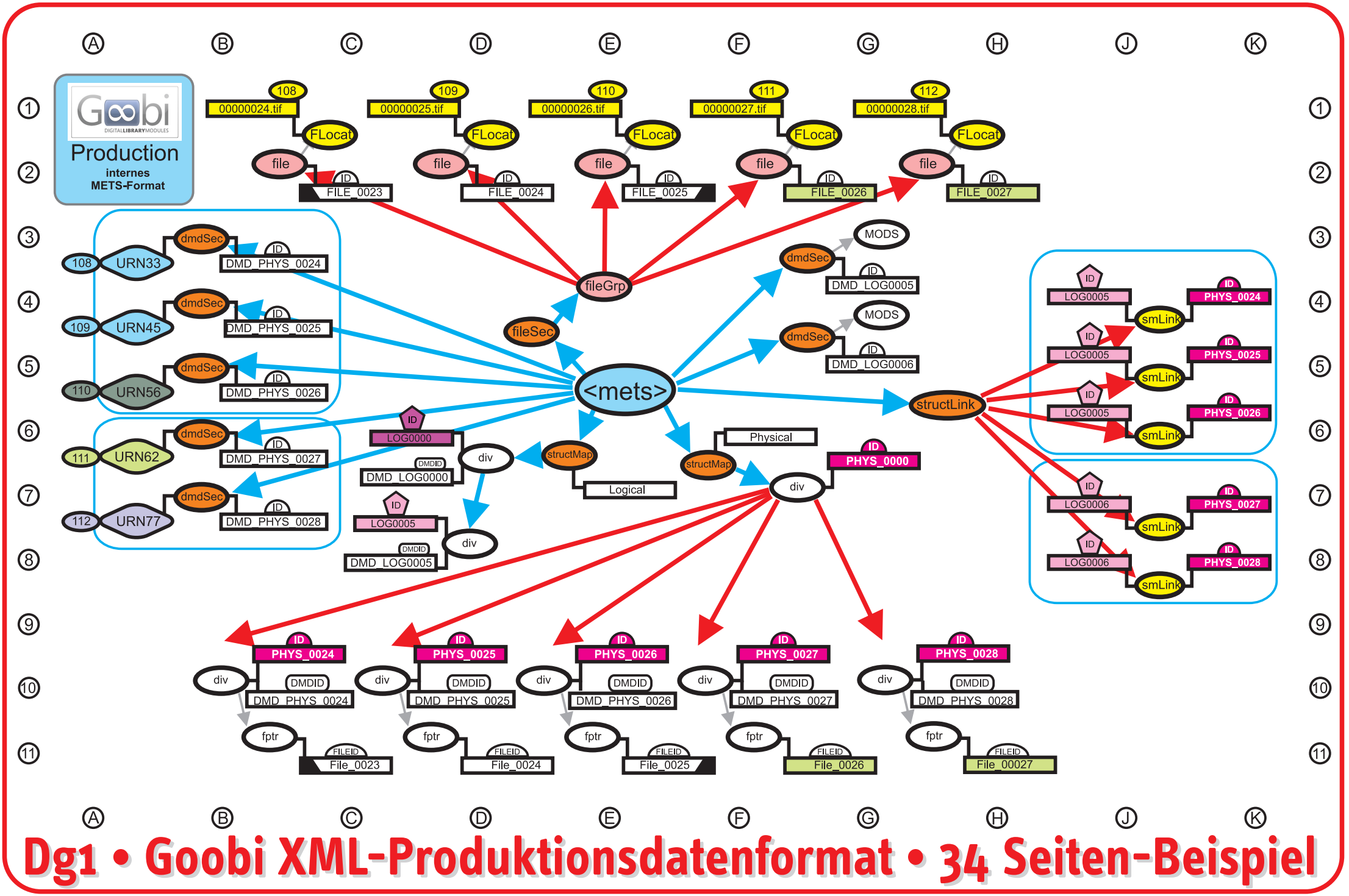
9.2 Erklärung zu Diagramm 1, 2 und 3

Die Diagramme zeigen nur beispielhaft Elemente zu den im Text behandelten Kapitelelementen. Diagramm 1 und Diagramm 2 zeigen das Produktionsdatenformat von Goobi. In Diagramm 1 wird die Situation mit 34 Seiten des Beispieldokuments gezeigt. Diagramm 2 zeigt die Situation nach dem Einfügen einer Bilddatei in den Bildstapel und den Versatz der URNs, wie in Kapitel 4.4. und 4.5 behandelt.

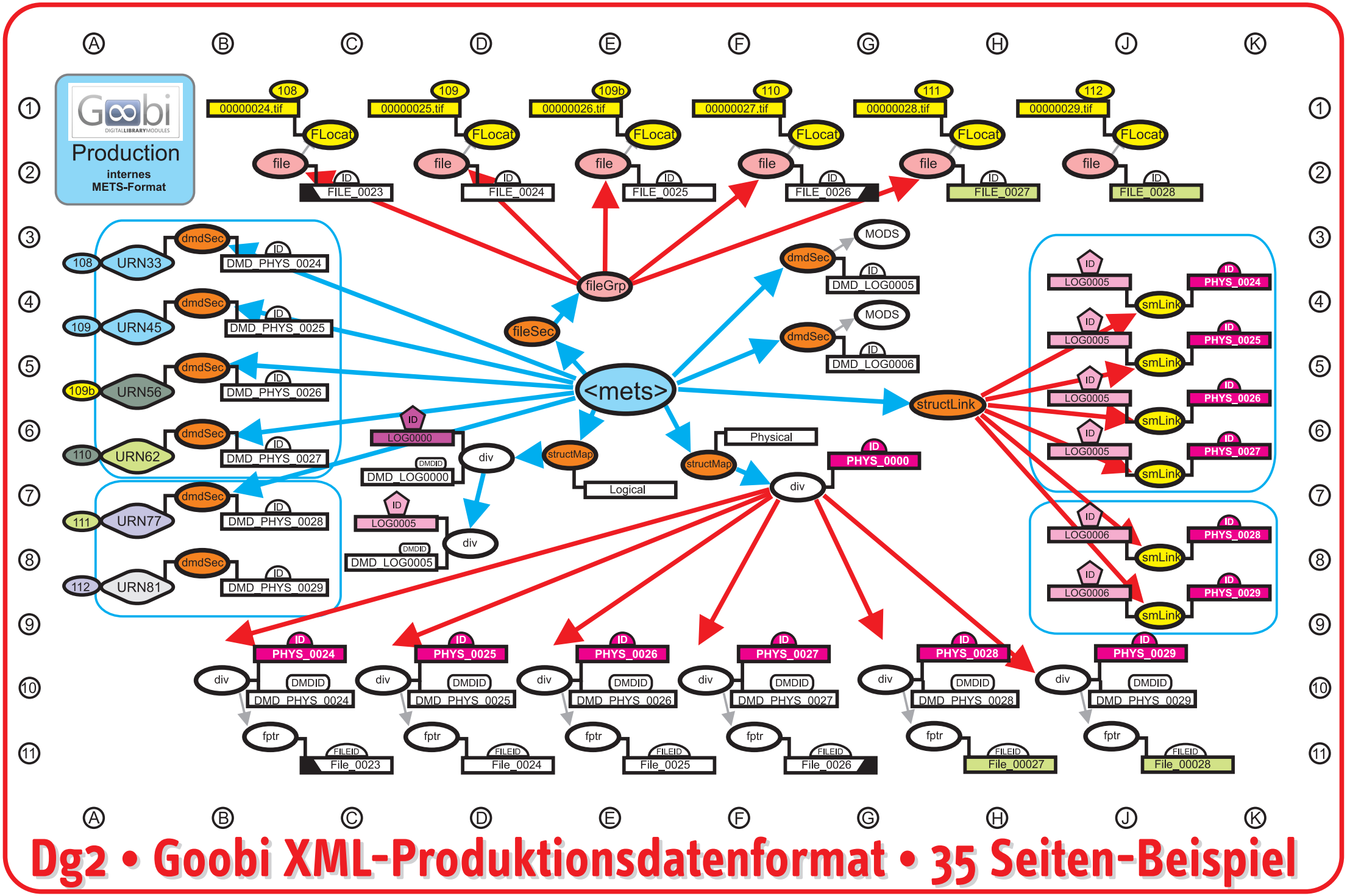
Das Diagramm 3 zeigt hingegen das valide METS-Datenformat nach der Korrektur des Bildstapels mit dem in Kauf genommenen URN-Versatz.

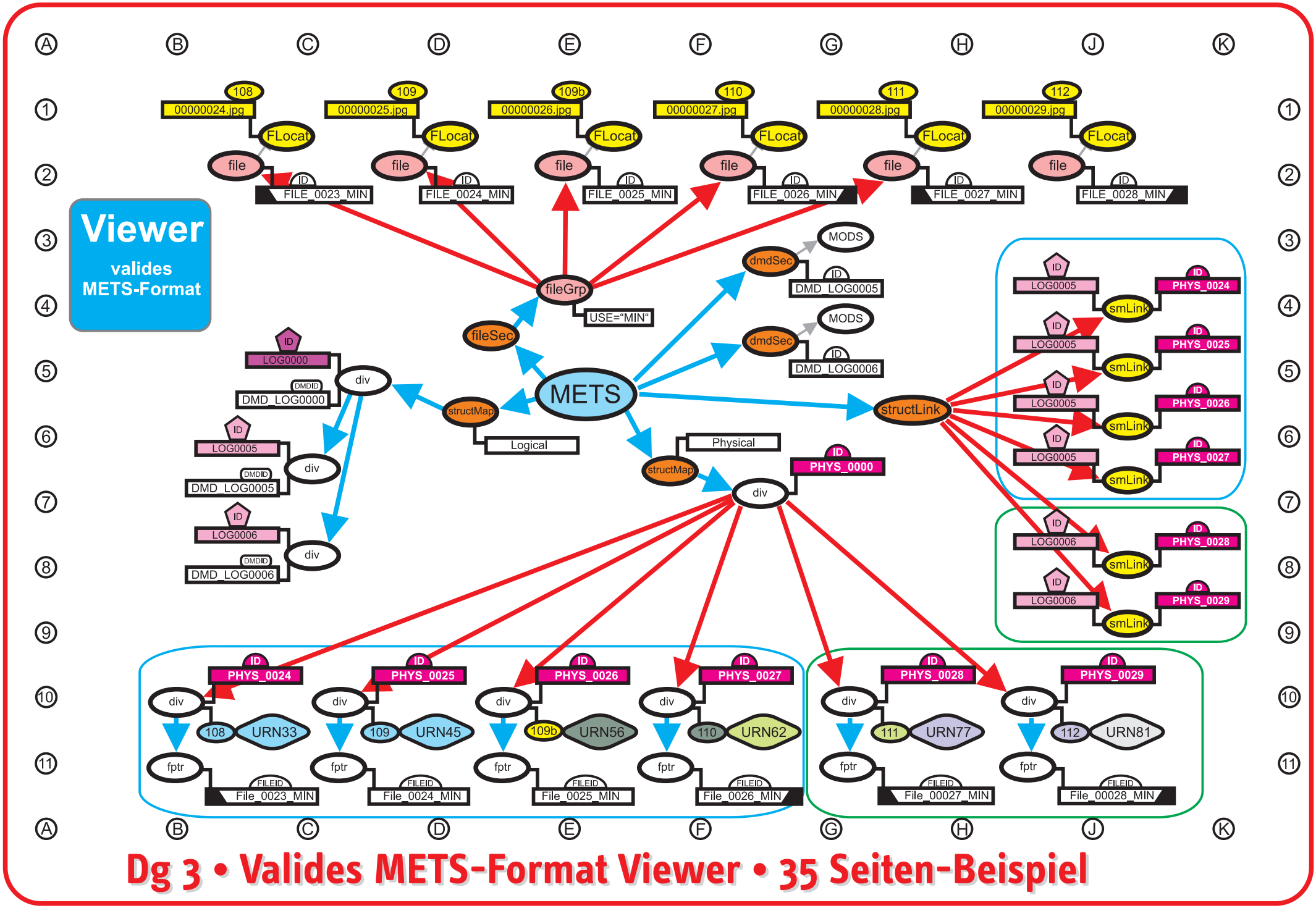


Dgo • Goobi XML-Produktionsdatenformat • Übersicht



Dg1 • Goobi XML-Produktionsdatenformat • 34 Seiten-Beispiel


Goobi
DIGITALLIBRARYMODULES
Production
internes
METS-Format
108
00000024.tif
109
00000025.tif
109b
00000026.tif
110
00000027.tif
111
00000028.tif
112
00000029.tif
FLocat
file
ID
FILE_0023
FILE_0024
FILE_0025
FILE_0026
FILE_0027
FILE_0028
fileGrp
fileSec
<mets>
dmdSec
MODS
DMD_LOG0005
DMD_LOG0006
structLink
smLink
LOG0005
LOG0006
PHYS_0024
PHYS_0025
PHYS_0026
PHYS_0027
PHYS_0028
PHYS_0029
108
URN33
109
URN45
109b
URN56
110
URN62
111
URN77
112
URN81
DMD_PHYS_0024
DMD_PHYS_0025
DMD_PHYS_0026
DMD_PHYS_0027
DMD_PHYS_0028
DMD_PHYS_0029
LOG0000
DMD_LOG0000
DMDID
div
structMap
Physical
Logical
PHYS_0000
fptr
FILEID
File_0023
File_0024
File_0025
File_0026
File_00027
File_00028
Dg2 • Goobi XML-Produktionsdatenformat • 35 Seiten-Beispiel



Dg 3 • Valides METS-Format Viewer • 35 Seiten-Beispiel